

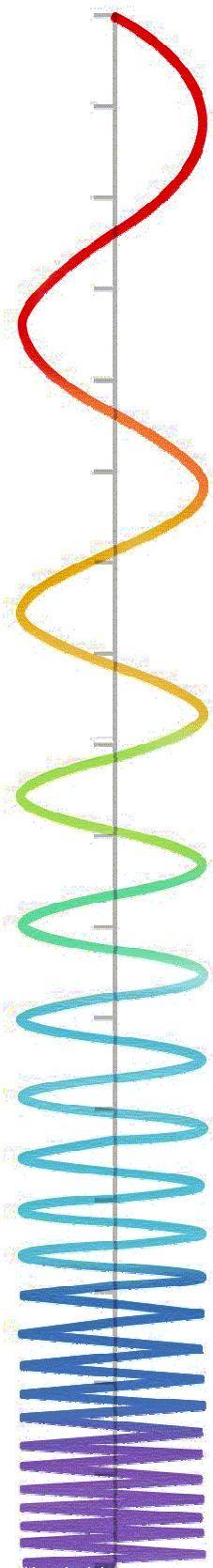
Quarterly Report
January-March 2003

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..... | 4 |
| 4 Systems Architecture | 9 |
| 5 Data Access/Resource Layer | 11 |
| 6 NVO Services | 14 |
| 7 Service/Data Provider Implementation and Integration | 16 |
| 8 Portals and Workbenches..... | 17 |
| 9 Test-Bed..... | 18 |
| 10 Science Prototypes..... | 18 |
| 11 Outreach and Education..... | 19 |
| Activities by Organization | 21 |
| Caltech–Astronomy Department | 21 |
| Caltech–Center for Advanced Computational Research | 21 |
| Caltech–Infrared Processing and Analysis Center | 22 |
| Canadian Astronomy Data Centre/Canadian Virtual Observatory Project..... | 22 |
| Carnegie-Mellon University/University of Pittsburgh..... | 22 |
| Fermi National Accelerator Laboratory..... | 23 |
| High Energy Astrophysics Science Archive Research Center | 23 |
| Johns Hopkins University | 23 |
| Microsoft Research | 24 |
| National Optical Astronomy Observatories..... | 24 |
| National Radio Astronomy Observatory | 25 |
| Raytheon Technical Services Company | 26 |
| San Diego Supercomputer Center..... | 26 |
| Smithsonian Astrophysical Observatory..... | 27 |
| Space Telescope Science Institute | 27 |
| United States Naval Observatory..... | 28 |
| University of Illinois Urbana-Champaign/ National Center for Supercomputer Applications..... | 28 |
| University of Pennsylvania..... | 29 |
| University of Southern California (ISI)..... | 30 |
| University of Wisconsin | 30 |
| Publications and Presentations | 31 |
| Acronyms | 32 |

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

Period covered by this report: 1 January-31 March 2003
Submitted by: Dr. Robert Hanisch (STScI), Project Manager

Executive Summary

Highlights:

- *Scientific.* Work is well underway to convert the gamma-ray burst follow-up service prototype into the NVO's first standard service, a Data Inventory Service. In addition, the new DIS will incorporate the project's first automated registries, allowing new resources to be added and discovered dynamically. Plans are being made for a science prototype that incorporates theoretical simulation data.

A press release was issued describing the discovery of a new brown dwarf from our brown dwarf science prototype demonstration.

- *Technical.* International technical collaboration has become very active since the January meeting of the International Virtual Observatory Alliance. Technical working groups have been established in six areas: registries, data models, UCDs (Uniform Content Descriptors), VO query language, data access layer, and VOTable.

NVO technical work has focused on registries and identifiers, and the OAI (Open Archives Initiative) mechanisms for managing registries have been studied and are being utilized in the development of a prototype. We are working toward having a basic registry service working prior to a May 2003 international technical meeting in Cambridge, UK. Progress was also made in refining the space-time metadata definitions and in specifying a syntax for describing regions.

An alternative naming convention for UCDs (Uniform Content Descriptors) has been developed, and will be discussed with our international partners at the May meeting in Cambridge.

Extensions have been proposed to the Simple Image Access Protocol, including adding support for spectral and time series data. Further discussions are needed, including international partners, on whether SIAP should be modified or whether separate protocols should be defined for other data types.

Metadata definitions for Education and Public Outreach products have been worked out and are being incorporated into the project's Resource and Service Metadata framework.

- *Programmatic.* Much work is going into preparing for the IAU General Assembly, with joint IVOA displays and demonstrations and Joint Discussion 8 on Future Large Telescopes and the Virtual Observatory. We held a team meeting in Pasadena, hosted by Caltech/CACR, in March.

Issues and Concerns: None at this time.

Activities by WBS

1 Management

1.1 Science Oversight (Executive Committee)

Status: We conducted a retrospective of the January 2003 science prototypes at our March 2003 team meeting. All agreed that such prototypes are a critical component of our project plan, as they force us to integrate and test components. We must reach a healthy balance, however, between making demonstrations and building core infrastructure. Thus, we decided to focus on updates to the January prototypes for the July 2003 IAU General Assembly, most notably with the incorporation of registry services into the gamma-ray burst follow-up service.

We also began planning for a science prototype based on a theoretical simulation data set, namely, a globular cluster simulation. This will be targeted for the January 2004 AAS meeting.

1.2 Technical Oversight (Executive Committee)

Highlights: In the International Virtual Observatory Alliance (IVOA) we reached agreement on the major technical focus areas for 2003. These include registries, data models, UCDS, data access layer, VO Query Language, and VOTable. Energetic collaborations have followed in the IVOA context, with a technical interoperability workshop being planned for Cambridge (UK) in May.

In this context, the NVO project began active development of resource and service registries, and participated in an IVOA technical discussion on this topic held in March.

Status: A software library has been established using the industry-standard CVS system, and is located at <http://nvo.gsfc.nasa.gov/viewcvs/viewcvs.cgi/>.

Mailing lists of general interest to the international VO collaborators have been consolidated at the International Virtual Observatory Alliance (IVOA) web site (<http://www.ivoa.net>). Many members of the NVO project are collaborating with the other VO initiatives through the IVOA TWiki (<http://www.ivoa.net/twiki/bin/view/IVOA/WebHome>). [A TWiki is a web-based environment for distributed project development; see <http://twiki.org> for more information.]

1.3 Project and Budget Oversight (Executive Committee)

Highlights: The project's Education and Outreach Coordinator, Mark Voit, will be leaving STScI this summer. We are transitioning the EPO responsibilities to Dr. Frank Summers, a scientist in the Office of Public Outreach at STScI with extensive experience in museum exhibits, planetariums, and data visualization.

Status: The project is now fully staffed and, with a few exceptions, spending at the expected rate. A few organizations are still very slow in submitting invoices, though no work is being held up. Project expenditures for this quarter were \$712,115, a figure that includes previously delayed invoices from several groups. See the financial supplement for additional information.

2 Data Models

2.1 Data Models / Data Model Architecture (McDowell, SAO)

Highlights: In this quarter we have begun an effort to define the connection between the existing UCD concept nametags and the attributes of the data model. A concept paper describing parameterized content descriptors and presenting an initial organization of the problem domain was circulated by J. McDowell on the IVOA mailing lists. Data modeling discussions have begun on the ivoa.net mailing list dm@ivoa.net.

Status: We are preparing for the IVOA meeting in England in May, where efforts will be made on the UCD problem and on data models for quantities and spectra.

2.2 Data Models / Data Types (McDowell, SAO)

Highlights: We are working on a spectral data model to support spectral energy distribution prototypes. Work with existing prototypes has drawn attention to the current lack of interoperable descriptions of the photometric calibration. J. McDowell met with the AVO/ESO team and was briefed on their data model for pipelines and observations.

Status: The goal is to generate an initial white paper on photometry models at the Cambridge meeting.

2.3 Data Models / Data Associations (McDowell, SAO)

Status: Work on generalized coverage (bandpass, regions etc) is ongoing;

3 Metadata Standards

3.1 Metadata Standards / Basic Profile Elements (Rots, SAO)

Highlights: Considerable progress has been made on spatial region definition. The re-write of the Space-Time Coordinate metadata specification, in the form of an XML schema, is essentially finished.

Issues and Concerns: For regions, the main issue identified is that we need to reconcile the needs of small-scale (e.g., source cut-outs) and large scale (e.g., survey coverage) region specifications. The underlying principles are clearly very similar, but the practical

requirements vary widely. Another area that requires attention is that of the various projections that are in use.

For Space-Time Coordinates, there has been a request to include all FITS WCS functionality; this is being studied at the moment and may well be included soon.

Status: The discussion of the spatial region definition has made considerable progress and we expect it to lead to a conclusion very soon. The new specification is more closely attuned to the way coordinates are usually handled and there have been considerable improvements in the XML implementation.

3.2 Specific Profile Implementations (McGlynn, HEASARC)

Status: Rather than the breakdown of types contemplated in the original proposal, it appears that profiles will be developed according to types of services. For example, ConeSearch and SIA services will have slightly different profiles. This is rather different than the ‘catalog’ versus archive profiles originally contemplated. There has been a lot of activity in this area during the past quarter motivated by the registry action. Effectively the profile defines the contents that are to be put in these registries. While there is considerable flux in these discussions, the May interoperability meeting (Cambridge, UK) should provide us with relatively detailed roadmaps for the further development of service-type profiles.

3.3 Metadata Representations and Encoding (Plante, UIUC/NCSA)

Highlights: In this last quarter, we have begun earnest development of a metadata definition framework based on the document “Resource and Service Metadata for the Virtual Observatory” (Hanisch et al. 2002, <http://bill.cacr.caltech.edu/cfdocs/usvo-pubs/files/ResourceServiceMetadataV6.pdf>) as a prototype (see WBS 3.4). We have focused on XML Schema as the machine-readable format for defining metadata and have successfully used XML tools to manipulate the metadata:

- NCSA (R. Plante and R. Williamson) has demonstrated the use of XSL to automatically produce a metadata dictionary on-the-fly from a schema, as well as convert VO RSM to Dublin Core.
- NCSA and Caltech (R. Williams) have demonstrated the use of VO RSM within an OAI interface (see WBS 3.6)
- STScI/JHU (G. Greene and W. O’Mullane) have demonstrated the use code-generating toolkits (e.g., Microsoft XSD) to automatically convert the schema into C# classes representing the RSM.

Currently, NCSA and STScI have been working together to discover good schema authoring styles that allow for clear metadata models, straightforward extensibility, and good compatibility with readily available XML tools.

As part of the general work on registry development (WBS 3.6), the Metadata Working Group also took up the issue of identifiers. A set of draft requirements was developed (<http://www.us-vo.org/metadata/>), and work was begun on a specification.

Issues and Concerns: Metadata schemas have close connections to data modeling and UCDs, which, from a management perspective, have separate development tracks. (In the NVO management plan, Data Models has a separate WBS; in the international forum, UCDs and data models have separate working groups.) Plante has pointed out the importance of these efforts being coordinated to avoid three separate standards for the same thing. This issue is expected to be discussed in the IVOA Interoperability meeting in May.

Still under study is how a metadata definition model affects one's ability to use existing XML tools for processing XML automatically based on the schema.

Status: With a stable version of the RSM in XML (by the May IVOA Interoperability Meeting), we plan to extend the discussion of these issues to the international VO forum. Work on an Identifiers standard will be resumed in the context of the IVOA Registry Working Group. A proposal for extending VOTable to allow extended metadata descriptions has been postponed in order to synchronize with the international VOTable WG roadmap, now under development.

3.4 Profile Applications (Plante, UIUC/NCSA)

Highlights: Raytheon TSC (E. Shaya and B. Thomas) has analyzed the use cases assembled in Year 1 and are designing a general architecture and set of XML-encoded languages for queries in VO applications. Separate languages have been designed for the different parts of the query processing chain. The high-level language allows users to express queries in terms most natural to them while at the same time allow query mediators to use registries to distribute queries to relevant repositories efficiently. The low-level language is simpler and is passed between a query mediator and searchable data repositories. All languages are metadata-neutral; that is, they allow one to form queries using any schema defined in XML, making the language extensible.

As part of a registry prototyping effort (WBS 3.6), NCSA (R. Plante and R. Williamson) have produced an XML Schema based on the RSM document. This effort has included a refinement of the Resource data model set down in that document which reconciles the concept of a Resource with that used in the general web community: that is, anything that is describable and identifiable can be a resource. Organizations, data collections, and services are different kinds of resources. The Resource metadata are then the descriptive concepts common to all resources. The XML Schema reflects this and allows for the data model to be extended to add concepts for the particular types of resources, most notably services. Refinement of the resource schema is focusing on specific support for describing Simple Image Access services, capturing the metadata specific to just this type of service (defined in the Simple Image Access Protocol Specification, Tody et al. 2002, <http://bill.cacr.caltech.edu/cfdocs/usvo-pubs/files/ACF8DE.pdf>).

The IVOA Registry working group has defined a work package dedicated to the registry metadata, which Plante is leading. In preparation for the May IVOA Interoperability Meeting, he will open up the discussion of the XML Schema for RSM to the international forum.

Issues and Concerns: At the moment, it is still unclear how the Query Language framework overlaps with the goals of the XML standard, XML Query, and whether it can be incorporated into the VO framework. We also need to better understand what work has been done in this area by other VO projects and to what extent we can integrate the efforts for an international standard.

It is expected that the XML-based modeling of the RSM will feedback suggested changes for the next version of the RSM document. Many issues are being left unsettled pending input from the international group. For example, some differences remain between the different VO projects on the form of resource identifiers; consensus on this issue will be needed before the schema can be finalized.

Status: E. Shaya and B. Thomas have generated a white paper describing their framework. A VO Query Language working group is now working under the auspices of the IVOA; representatives from the NVO are participating. Resource metadata will be discussed at the May IVOA Interoperability Meeting in the context of the Metadata Specifications work package of the Registry Working Group.

3.5 Metadata Standards / Relationships (Rots, SAO)

Highlights: We have drafted a document entitled “Outreach Metadata for the Virtual Observatory” describing the E/PO metadata needs for NVO.

Status: The E/PO metadata are now being incorporated in the project-wide metadata definitions. We are documenting specific prototype examples showing how NVO metadata will be used to describe education and outreach materials, so that our partners and other developers can learn how to apply these concepts.

3.6 Metadata APIs (Plante, UIUC/NCSA)

Highlights: R. Plante has started a draft specification for VO identifiers that defines services used for resolving identifiers into metadata descriptions of resources (see link at <http://www.us-vo.org/metadata>).

In our evaluation of the first year science prototypes, the incorporation of registries was identified as the most important missing component needed to generalize the demos for production use by real astronomers. Thus, the Metadata Working Group (MWG) initiated a rapid development plan for NVO registries with a three-step process:

1. Assembling requirements and use cases for registries.
2. Drafting a rough specification for registries.

3. Prototyping the registry specification.

Requirements and use cases were developed and assembled on the MWG page (<http://www.us-vo.org/metadata>). Initial prototyping proceeded ahead of specifications as a way to explore existing technologies. As mentioned in WBS 3.4, a draft specification for resource identifiers, a critical component of a registry, was started by Plante.

One of the technologies explored in detail was the Protocol for Metadata Harvesting (PMH) developed as part of the Open Archives Initiative (<http://www.openarchives.org>). We consulted with the protocol's primary author, C. Lagoze (also a member of the External Advisory Committee), to explore ways that it might be used as part of an VO registry framework. It was identified as a potentially useful technology for transmitting resource descriptions from data providers to queryable registries. Several groups developed prototype implementations of the protocol:

- STScI/JHU (G. Greene and W. O'Mullane) explored a web services variation on the protocol.
- Caltech (R. Williams) assembled a registry of Simple Image Access services and exposed them via an OAI interface.
- NCSA (R. Plante and R. Williamson) prototyped a deployable, CGI-based OAI implementation aimed providing a low-cost way for repositories to describe themselves.

All three groups have benefited from trading various technology solutions between them.

A Registry Prototyping "Tiger Team" was assembled (including NCSA, Caltech, STScI, and HEASARC) to connect the various registry components under development into a prototype registry that could be used with the Data Inventory Service (the successor to the gamma-ray burst science prototype) in time for the IAU General Assembly in July.

Meanwhile, IVOA sponsored a joint planning meeting for registries in March. Various representatives from the NVO attended by telecon. This meeting marked the official "kick-off" of the IVOA Registry Working Group, leading to the definition of five work packages that define an internationally coordinated roadmap for the specification and development of registries. NVO is actively participating in this work; in particular, R. Hanisch is assisting with the development of documentation standards and policies, and R. Plante is leading the Metadata Specifications work package.

Issues and Concerns: The creation of the IVOA working groups will necessitate some rescheduling of our development activities to synchronize with the international efforts.

It is yet unclear exactly how OAI should be deployed for collecting metadata into a registry. In particular, there are the questions of who (e.g., data providers, metadata brokers/proxies) should set up an OAI service and under what circumstances. Also, details regarding the form of the metadata being exposed via OAI are still under discussion.

Status: A prototype registry for SIA and ConeSearch services should be ready for use by the Data Inventory Service by the end of spring 2003. Various open questions regarding

registries, resource metadata, and resource identifiers will be discussed at the May IVOA Interoperability Meeting.

4 Systems Architecture

4.1 System Design (Moore, SDSC)

Highlights: The architecture that integrates web-based services with grid technology is still being revised, based upon practical experience learned with NVO demonstrations, the NSF Teragrid, and support for NVO collections. The grid technology is still evolving too rapidly to propose a fixed implementation. We do know that many of the capabilities that are desired for the NVO can be met with grid technology. The architecture will include portals, web services, data access layer, collection management, and grid services.

The web services design has been primarily led by D. Tody and R. Williams. The data analysis support will be provided by porting data processing pipelines onto the Globus toolkit and the Chimera system, led by E. Deelman and J. Good. The Data Grid support is being demonstrated through the Storage Resource Broker, led by R. Moore.

Issues and Concerns: 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.

The implementation of grid services based on the emerging OGSA standard still needs practical experience to drive the design. Multiple groups are implementing WSDL-based basic data access services that can then be tested for the required performance levels.

Status:

4.1.1 System Design. The system design of the NVO architecture has the following components:

1. Portals – web service interfaces to analysis procedures (OASIS)
2. Process management systems – data processing pipelines to create derived data products (Chimera, Montage)
3. Web services – uniform capabilities provided across NVO catalogs and image archives (ConeSearch, VOTable catalog query, simple image access)
4. Data access layer – management of methods on data encoding formats for access based on physical quantities (UCDs)
5. Data grid – management of distributed collections, and provision of logical name space for global persistent identifiers (SRB)

6. Computational grid – access to distributed compute resources (Globus toolkit)
7. Persistent archives – management of technology evolution (SRB)
8. Astrophysics catalogs and Image archives (SDSS, 2MASS, DPOSS, USNO-B, MACHO)
9. Persistent disk systems – interactive access to sky survey image collections (Grid Bricks)
10. High performance disk caches – high speed access for bulk data analysis (SAN)
11. Compute platforms – NSF Teragrid

The architecture specifies seven software layers and four resource layers. Components have been developed for all of the layers, with the exception of the data access layer. 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 of the NVO architecture.

The area that is least well defined is the data access layer. This 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 Uniform Content Descriptors, and the ability to organize semantic terms relative to the NVO concept spaces for space, time, and domain knowledge. Once the data model is specified, it will be reasonable to implement a data access layer.

4.1.2 System-Level Requirements Definition, and 4.1.3 Interaction with Grid Components and Tools. The Grid architecture continues to evolve, under pressure from multiple application areas, including the NVO. An example of an emerging NVO requirement is the implementation of high performance image cutout services. In the Galactic Morphology demo, the images were moved under control of the Chimera system to a compute platform where they were processed. In the production DPOSS image access system, the image cutouts are generated directly at the storage system, eliminating the need to move GB-sized images over the network. A combination of these mechanisms is needed to improve performance of NVO services. This in turn means that distribution of computation within the NVO testbed must span both the data grid management layer and the computational grid execution layer.

A second system-level requirement is the development of a common web-service interface for all NVO functionality. The Open Grid Service Architecture WSDL interfaces are still under development for both the Globus Toolkit and the Storage Resource Broker data grid. For the SRB, WSDL-based data access services are available.

4.1.4 Logical Name Space. A current debate within the NVO is on the differentiation between identifiers that define uniqueness (Object Identifiers – OIDs), identifiers that define location (logical name space or replica catalog), and identifiers that define physical files (physical file name). The registration of unique OIDs will require

development of publication mechanisms typically associated with digital libraries. The usage scenario would be:

- Based on an OID determine which collection has the requested image
 - Ask the collection to determine the logical name (multiple possible storage locations)
 - Map from the logical name to physical file name used at the remote storage repository
- Note that several collections could hold a requested image, each collection using a different logical name for the image, with the images stored at multiple sites.

A second discussion point is the mapping of UCDs to the logical name space. Through the data access layer, one would like to make requests based upon UCDs, without having to know the logical name for an image. In practice, the logical name space of a collection is used to represent all registered digital entities. State information associated with NVO services are mapped onto the logical name space. The state information can be stored in multiple registries (with the logical name used as the foreign key to do joins across the registries), or the state information can be stored in a single catalog. The use of the logical name space to manage NVO service state information highlights the need for a level of naming indirection between uniqueness identifiers and the logical name space.

4.2 Interface Definition (Williams, CACR)

Status: Discussions are continuing on possible extensions to the Simple Image Access Protocol. SAO has been working on a requirements and specification document for extending SIAP; there are a number of archives and data depositories that are not served well by the current protocol. We are also working with our international partners to reach agreement on the SIA definition for all VO projects.

4.3 Network Requirements (Williams, CACR)

Status: Work not scheduled until late in CY2002.

4.4 Computational Requirements (Williams, CACR)

Status: Work not scheduled until late in CY2002.

4.5 Security Requirements (Deelman, USC)

Status: No work done in this quarter. The AstroGrid project in the UK is actively investigating authentication and resource allocation issues in the Grid framework, and we are likely to follow their lead in this area.

5 Data Access/Resource Layer

5.1 Resource and Information Discovery (Szalay, JHU)

Status: Work in this area has focused on resource and service registries (described in WBS 3.6).

5.2 Data Access Mechanisms (Deelman, USC)

Status. USC/ISI is evaluating the use of the Globus Replica Location Service (RLS) in the context of Chimera/Pegasus when used in the context of the galaxy morphology science prototype and the Montage application. The RLS is used by Pegasus to locate the copies of the existing data. Based on the availability of the data, Pegasus is able to reduce the abstract workflow described by Chimera to its minimum computational jobs. For example, if data is available on the Grid, Pegasus will reuse that data rather than recomputed the data products. The RLS is also used to find the locations of the necessary input files. When the computation is completed, RLS is used to register the newly derived data products for future use.

5.3 Data Access Protocols (Williams, CACR)

Status: The NVO Simple Image Access Protocol is being utilized in the grid-based version of Montage (Montage-G) for data access. SIAP will be used to transfer thousands of image files representing many terabytes of data to the NSF Teragrid. Caltech has built an SIAP service for the pre-release 2MASS dataset, so that Montage-G can be used on this dataset. The SIAP has switches to allow data to be fetched from either SDSC or Caltech. Montage-G 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.

In order to better support large-scale applications, such as those targeted by NVO, additional features needed to be included in the Pegasus software. USC/ISI has been working on adding such features as the transient file features and read modes to enable Pegasus to efficiently handle workflows with large numbers of nodes. The description of these features follows:

- 1) Transient File Features. Pegasus now supports transient files, which gives the user the facility to specify which particular files he wants to be transferred to the output pool and registered into the RLS. This feature can potentially drastically reduce the load on the job managers of the pools in case of very large DAGs by not doing unnecessary transfers of data products.
- 2) Read Modes. Pegasus now supports two read modes for the `pool.config` file and `tc.data` file. The single read mode ends up reading the files into memory on startup and the multiple read mode parses the file as when required. The single read mode is default and expected to be beneficial for most users compared to other approach. The multiple read mode has been kept to handle very large data files, which may exceed the system's memory limitations and make it impossible to load it completely into memory.

5.4 Data Access Portals (Tody, NRAO)

Highlights. With the completion of the data access service prototypes (e.g., Simple Image Access) and their successful use in the VO science application prototypes demonstrated at the AAS in January, attention has turned to planning the second phase of VO data access layer (DAL) development. This will include an expanded set of standard data access services, integration with registries and data models, and a scalable framework for grid-enabled data access and analysis. Preparations are underway for the IVOA interoperability workshop to be held in Cambridge, UK in May. An IVOA data access layer working group has been formed to define international standards for VO data access.

Status. The concept for the Data Access Portal and Generic Client interface presented in the NVO management plan describes a phased approach with the following as the principal deliverables of the first two phases of development:

- A prototype providing basic Web-oriented data access capabilities based on URLs and FITS+XML to support intermediate science scenarios (July 2003).
- An expanded client interface and set of data access services and protocols, and a distributed, scalable data access / analysis framework implementing the VO data models, adding support for data mediation and server-side (distributed) computation, including application of analysis or transformation functions for virtual data generation, and supporting integration with legacy systems (January 2006).

The prototype facilities outlined for Phase 1 are the existing simple image access (SIA) and ConeSearch services. These have been in use since late 2002, and are being enhanced for the next round of science prototypes, e.g., by adding support for dynamic registries and by further extending the image metadata and data model. A call for proposals for version 2 of SIA has been issued, including a summary of areas where the standard could be extended. A second version of SIA is planned for this summer.

Much of the effort this past quarter has gone into planning DAL phase 2. Phase 2 integrates the prototype facilities and other VO technology into a more general Grid-capable, distributed, scalable data analysis framework. The existing ConeSearch and simple image access services will evolve into a collection of services for accessing the full range of data, e.g., catalogs, images, spectra, time-series data, and so forth. Standards now being developed for registry integration, queries, metadata, and data models, will be integrated into the new data access service protocols. These protocols will be defined by a data access layer working group now being formed within the IVOA.

In addition to an expanded range of data access protocols, a scalable data analysis framework is needed to integrate conventional astronomical data analysis and VO, and provide a scalable framework for implementing VO services and distributed data analysis. Much of the effort this past quarter has gone into research to develop the concepts for such a new data analysis framework. A draft whitepaper exploring the

framework problem was released internally within the NVO in early April. This will be developed further over the next quarter.

6 NVO Services

6.1 Computational Services (Berriman, IRSA)

Highlights: The Montage team has been performing formal testing of the Montage compute engine under the aegis of the NASA ESTO-CT program. The operational version of Montage will run on the emerging Distributed Terascale Facility (the *Teragrid*). We have begun collaborations with USC/ISI for running Montage in a grid environment. Our goal is to understand which technologies provide the most stable environment on which to run Montage, and thereby design and deploy a “grid-enabled” Montage. USC/ISI has initiated the port of the Montage application to the Chimera/Pegasus environment. The Montage version used for the computation was 1.4.

Status: In order to run Montage in the Chimera/Pegasus environment ISI generated the equivalent transformations and derivations for all the Montage methods used. Output and input arguments such as image tables, correction tables, etc., were given logical file names that are used by Chimera to identify dependencies between the operations. These logical file names are the parameters in the Chimera derivations. An abstract directed acyclic graph (DAG) is created for computing the final image mosaic FITS file. The Pegasus planner transforms this abstract graph of derivations produced by Chimera into an executable DAG. This executable DAG is then submitted to Condor-G to be executed over the Grid.

To evaluate the Chimera/Pegasus-based Montage, the application was run on a pool of Linux machines at ISI. The input to the computation was a cache of 2 Micron All Sky Server (2MASS) image files in FITS format and the output is the final image mosaic FITS file. Montage uses 12 distinct operations to achieve the final output (mImgtbl → mMakehdr → mSubset → mProjExec → mImgtbl → mOverlaps → mDiffExec → mFitExec → mBgModel → mBgExec → mImgtbl → mAdd). There were 91 image files in the input set. The number was chosen keeping memory requirements in mind. The computation ran successfully and completed in approximately 5 hours.

Request Management. The last quarterly report announced the development of version 1.0 of the Request Object Management Environment (ROME). In this quarter, IPAC has investigated how applications can interface with ROME seamlessly and easily.

A CGI program can easily be modified to become ROME compliant by adding the following components:

- *Keyword Inputs:* If the application is started by ROME processor, it should include the inputs mode=rome and ws=workspace_name in the request string.

- *Environment Paths:* An application program needs to find resources such as system utilities, various server programs, and the DBMS tables etc. These environment paths are normally set by the Apache server configuration file. When the application is started by a ROME processor, these properties needs to be set by the application program, since by definition, ROME does not talk to Apache.
- *Message Handler:* A ROME application should output messages via the standard output pipe to report then job status; at the very least it should report the status when the job started and ended. When the job is successfully started, a ROME-compliant application should output a message containing the job ID and the location of the message HTML file (in the “workspace” or staging area) in the appropriate XML format. This message will be written to the ROME request table so that when a client wishes to abort a job, the job ID can be retrieved and used to abort the job; similarly when a client wishes to view the full messages reported for this job, the message HTML file can be retrieved. When the job is completed, a ROME-compliant application should output a message that includes the data URL so that data can be retrieved.
- *Signal Handler:* When there is error condition, a ROME application should trap the error signal and passes the error message to ROME and the client.

IPAC has developed a set of C library routines to be used with CGI programs so that the same application program can serve both as the CGI program and a ROME application program. Furthermore, to add the above necessary components to a CGI program, an application only needs to call a simple C routine rather than to insert a block of codes for constructing the XML messages. The following sample C code sets the environment, the signal handler, the message handler, and to output a message.

```

if (isrome) {
    rome_setenv();
    rome_sigset();

    rome_msginit (directory, msgfile, server, urlbase,
                 appname, process_id);
    rome_setmsg ("start", msg);
}

```

We have begun the development of a ROME testbed at IPAC, but this effort has been delayed because the developer has taken extended disability leave.

6.2 Computational Resource Management (Moore/SDSC)

Issues and Concerns. The set of resources that will be manipulated within the NVO testbed are evolving to include:

- Image archives
- Image disk-based collections
- Catalogs
- Compute resources

Each of these resources has an associated computational requirement, whether the creation of image cutouts directly at the image archive, or the processing of complex queries against the catalog, or the execution of applications on remote compute platform. The software systems that provide these capabilities are:

- DataCutter data filtering system, SRB remote proxies
- OGSA-DAIS data access and information system
- Chimera, Globus toolkit

While we have understood all along that we needed to manage computation, we have not understood the implications of managing data filtering and large-scale queries. The impact is that the NVO testbed will need to manage more sophisticated sets of operations than most other grid environments. The closest similar system is a DOE SciDAC environment that supports complex queries on bio-informatics databases.

The major impact on the NVO system design is the development of support for large-scale queries on collections. Should this be managed through data mediation (mapping of concept spaces onto the NVO collections), through union catalogs, or through dynamic implementation of a query space on which the joins are carried out? Experience currently indicates that doing the dynamic join across catalogs is too inefficient, unless bulk metadata transport mechanisms can be implemented.

7 Service/Data Provider Implementation and Integration

7.1 Service/Data Provider Implementation (Hanisch/STScI)

Status: The ConeSearch and Simple Image Access Protocol services will be the first entries in the initial service registry. We have begun development of the registry (see WBS 3.6), and this will include interfaces that allow data and service providers to easily create and update registry entries.

7.2 Service/Data Provider Integration (Hanisch/STScI)

Status: Participating organizations continue to implement Simple Image Access Protocol-compliant services. IPAC is supporting the development of SIAP and VOTable compliant services for serving NED data. This work has involved re-organization of the NED database tables, brought about because many of the images were not “http accessible” and many do not have fully formed FITS headers, and those that do have FITS headers do not have information on the image parameters available in a searchable form.

IPAC created and populated new Informix database tables that contain object names and coordinates (right ascension and declination in J2000 for the corners of the image), image parameters from the FITS headers (as available), bibcode of paper (as available), date of observation in free format; and a WCS quality flag q_{WCS} to indicate whether the images are FITS or JPG format and if the WCS information in the header is usable in searches for images (“usable” here means that the footprint and location of the image on the sky are fully described).

- `qwcs=11` stands for FITS images with usable headers (cutouts from the DSS survey, which are archived in NED, fall into this category)
- `qwcs=10` stands for FITS images with invalid FITS/WCS information
- `qwcs=22` stands for 2MASS Survey images, for which NED archives JPG images and links to the IRSA Archive for accessing archived data
- `qwcs=21` stands for JPG formatted images of radio maps, graphs, and images from old atlases (like Carnegie or Zwicky); these do not have four corners or coverage information.

IPAC has used these new tables to return output from SIAP- and VOTable-compliant services, currently under test.

IPAC has also investigated how to deploy NVO-compliant services through IRSA in a fashion that involves minimal impact on the IRSA architecture. Rather than deploy separate versions of each service, we have settled on a design that involves having a single version of each service, but this version supports several “modes” of operation, including NVO compliant output, with standardized input and output parameters.

8 Portals and Workbenches

8.1 Data Location Services (McGlynn, USRA/HEASARC)

Highlights: Major activity continued in this area. Starting with the GRB demonstration service, a region inventory service is being developed which provides a summary of information available for any service. Release of this service to the public is anticipated in Q2 CY2003. There are several substantial changes to the original GRB service. The user interface has been simplified. Target names and positions are both supported. There is active management of the files produced. Identical requests are not reprocessed but are pointed to the results for the original request. Some work remains in going through the list of services and ensuring that they are all being queried correctly. Some new services can be added.

Issues and Concerns: Initial public release may either use an internal registry of services or may be delayed until it uses a remote queryable registry. Currently several registry systems are being developed at both US and IVOA sites. The choice of the registry to be used for this service has not yet been made.

Status: The development of a replacement for Astrobrowse is largely complete. The SkyView SIA protocol is in place and provides integration of most SkyView surveys within the VO.

8.2 Cross-Correlation Services (Djorgovski, Caltech)

Status: No scheduled activities prior to CY2002 Q3.

8.3 Visualization Services (Williams, CACR)

Status: No scheduled activities prior to CY2003 Q3.

8.4 Theoretical Models (De Young, NOAO)

Status: Discussions were initiated to define a science prototype based on theoretical simulations. See WBS 10.2.

9 Test-Bed

Status: FNAL, USC/ISI, and NCSA are currently building 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 issue that needs to be addressed in the FNAL environment is the use of Kerberos certificates. ISI is currently working on a solution to the problem.

We have planned to test software on the NVO testbed, which is slowly coming up to production status on top of the NSF Teragrid. When the NSF Teragrid is in production operation, we expect to evaluate the collection management software, the grid software, and the OGSA services software. The NSF Teragrid is now expected to run in 2nd quarter CY2003.

10 Science Prototypes

10.1 Definition of Essential Astronomical Services (Szalay, JHU)

Status: Technical discussions continued with the AVO, AstroGrid, and other international partners in defining essential VO services. VO Japan has agreed to take the lead in the definition of a VO Query Language, and we are working jointly on service registries, refinement of UCDs (Uniform Content Descriptors), and Web Services.

10.2 Definition of Representative Query Cases (De Young, NOAO)

Status: Discussions were initiated with NVO partners interested in theoretical models and simulations to see what science prototypes might be implemented in the coming six to nine months. P. Teuben (U. Maryland) visited STScI/JHU to discuss such possibilities in detail, and we are currently planning a demonstration project based on a globular cluster simulation data set. The prototype will address the question of mass segregation in globular clusters as ascertained from making simulated observations of a suite of models, and then comparing their “observed” properties to actual observations from HST, Chandra, and other observatories. Planning is underway to show this demonstration at the January 2004 AAS meeting.

10.3 Design, Definition, and Demonstration of Science Capabilities (De Young, NOAO)

Highlights: The continued development of the gamma-ray burst follow-up prototype remained a key focus of activities during the past quarter. The service was successfully demonstrated by many sites at the Seattle AAS meeting. The service was generally found to be useful and feedback indicated that transitioning this service to kind of region inventory service independent of any link to discrete temporal events would be well received. Work is now in progress to provide this as the first general NVO service, rechristened as the Data Inventory Service. Plans for the demonstrations at the July IAU meeting are to enhance the service to allow use of a dynamic queryable registry rather than the internal static list of services used in Seattle.

Status: The original prototype remains fully operational. A distinct and simpler interface to the same basic service has been developed where the user gives only the position and size of the region of interest. See WBS 8.1 for more details. An enhanced version of the prototype that uses dynamic registries will be presented at the July IAU meeting.

11 Outreach and Education

11.1 Strategic Partnerships (Voit, STScI)

The Hands-On Universe team has proposed an NVO-based education project to the NSF Math and Science Partnership program. The project would create “collaboratories” that bring students and teachers together with scientists, educators, and programmers to use astronomical data through the NVO. Initially the project would focus primarily on the SDSS dataset, and the Hands-on Universe project is already working in partnership with the SDSS outreach lead. (This collaboration began at the July 2002 NVO Outreach Workshop.)

11.2 Education Initiatives (Voit, STScI)

No activities to report this Quarter.

11.3 Outreach and Press Activities (Voit, STScI)

Highlights: We produced a press release describing the discovery of a brown dwarf in one of the science prototypes developed for the January 2003 AAS meeting: “Virtual Observatory Prototype Produces Surprise Discovery,” *Headlines@Hopkins*, 12 March 2003, http://www.jhu.edu/news_info/news/home03/mar03/nvo.html. Follow-up articles appeared in Spaceflight Now (3/12/03), SpaceNews International, and NPACI Online, and UPI released a story (also on 3/12/03). The New York Times, Der Spiegel, and the IEEE are preparing stories.

Status: The development of the release raised issues about how to properly acknowledge all institutions involved in NVO-enabled discoveries. Efforts are underway to craft a policy addressing these issues.

Popular Press Articles About NVO and IVOA:

“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

“European Virtual Observatory One Step Nearer,” *Sky and Telescope*, 28 January 2003, 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 (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

Activities by Organization

Caltech—Astronomy Department

A. Mahabal has been working on a topic map made from observing logs. Such a topic map can be generated per telescope per instrument. These topic maps can then be combined (merged) to allow users to ask various questions. Example questions just concerning the objects are:

- (1) What telescopes was object XYZ observed with?
- (2) What total exposures have been reached for a certain object?
- (3) What dates was a particular object observed on?

A topic map based on data from the Palomar Large Format Camera can be seen at http://avyakta.caltech.edu:8096/omnigator/models/topicmap_complete.jsp?tm=lfcoobslogtm.xtm.

Caltech—Center for Advanced Computational Research

At Caltech, a postdoc (M. Graham) has been hired to begin work on July 7, who will be half time NVO and half time on the Quest synoptic sky survey. We expect that the data processing and dissemination of Quest results will be strongly based on NVO protocols.

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 Cambridge, where we expect to solidify the meaning and syntax of UCDs.

The document on NVO Resource and Service Metadata that was presented and solidified in 2001Q4 has been formalized to an XML schema (VOResource.xsd), so that registries can exchange and present these records. We expect in the next quarter to have a query service acting on these collections of records.

The related NASA-VO project, Montage, has advanced almost to the point of public software release (expected June 03). R. Williams is a member of this team, and has been participating in both the software engineering, and on building a Grid-capable version of Montage, called Montage-G.

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.

Caltech has begun the “hyperatlas” project, beginning with development of a naming scheme for coherent collections of FITS headers. Such a collection has been called an

"atlas", and each FITS header is a "chart" upon which image data can be projected and mosaiced. The intent of the hyperatlas project is to encourage image resources to be projected to the same pixel grid, so that sophisticated data mining algorithms can be brought to bear on the pixels. 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.

Caltech–Infrared Processing and Analysis Center

During this period IPAC

- Designed the architecture for applications to interface with ROME
- Cooperated with other NVO staff to deploy Montage on computational grids, especially the Teragrid
- Developed database architecture to support NVO compliant services through NED
- Developed architecture to support NVO compliance of IRSA

Canadian Astronomy Data Centre/Canadian Virtual Observatory Project

The CVO Database Exploration prototype was released to the public in February following internal testing and debugging. The deployment is currently on Sybase and has serious performance issues, forcing us to place limitations on the allowable queries and downloads.

Purchasing and installation of a 16 processor 7-Terabyte IBM database was completed in March 2003. Configuration and performance optimization is currently underway with public release of the CVO prototype system on IBM hardware is expected in June 2003. Demonstration of functionality will be done at the IAU General Assembly.

CADC is preparing the production of the Phase II content for the WFPC2 Associations stacks in collaboration with STScI and ST-ECF. These stacks form part of the pixel and catalog content of the CVO prototype. Other content development work involves the CFH12k camera, MegaPrime, and ACS on HST.

A major accomplishment has been the development of a query data model that is generic and designed for multi-wavelength content. We are collaborating with the German VO group (GAVO) and the Australian VO group in order to incorporate X-ray observations and catalogs and spectroscopic observations into this generic data model thus testing its validity.

Carnegie-Mellon University/University of Pittsburgh

We continue to develop and document our fast and efficient statistical algorithms as part of the PiCA group. For example, we have now completed extensive internal and external tests of our fast n-point correlation function code, including Bob Nichol (CMU) running the code on a mock SDSS catalog constructed from the Hubble Volume simulation. We have also begun to develop an approximate method for the n-point code that drastically reduces the computational time of this algorithm. Jeff Gardner (UPitt) presented results at the latest NVO meeting about parallelizing this n-point code. Jeorg Colberg (UPitt) continues to develop a web service in .NET for the n-point code that will result in an NVO-compliant version of that code. In summary, our ultimate goal is to provide the

NVO with a tested and documented version of the correlation function code, which can be accessed as a web service or requested as a GRID (parallelized) application. This will be our prototype NVO algorithm. Andrew Moore (CMU) continues to make progress on getting all our software into a simple-to-access table-based interface.

Fermi National Accelerator Laboratory

J. Annis attended the AAS meeting in January and provided support for the cluster morphology science demo (WBS 10.3.1).

Major effort has been spent preparing the first SDSS data release (DR1), which will supersede the EDR (early data release). It is expected that DR1 will be a key component of the NVO. Vijay Sekhri continued to develop a Simple Image Access Protocol interface to the SDSS image cutouts, adding WCS coordinates to the image headers (WBS 7.1).

Vijay Sekhri worked on a 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 could be useful for integrating NVO with grid computing resources (WBS 5.2.2).

High Energy Astrophysics Science Archive Research Center

The primary focus of the HEASARC during the first quarter has been the continued development of the gamma-ray burst prototype and its transition into a region inventory service. HEASARC personnel supported the demonstrations of the GRB prototype at the Seattle AAS meeting where it generally received good response. Since then, a variety of changes and enhancements have been made to make the service more generally useful, making it a data inventory service. The service has been completely transferred to the HEASARC's web service development machine in preparation for its release. A presentation describing the lessons learned in building the GRB prototypes was submitted to the NVO team meeting.

Other major activities include continued participation in regular metadata and management telecons, and active involvement in the discussions of metadata, registries and services in both the VO and IVOA forums.

USRA hired Dr. Jeongin Lee at the end of February to provide major elements of the HEASARC support for NVO activities. Dr. Lee is now responsible for the GRB/ data inventory service development.

Johns Hopkins University

The HTM website has been tidied up. HTM is a prospective target platform for footprint/region work for NVO. We are currently working with A.Rots(SAO) to define the region specification; considerable progress was made at the Pasadena team meeting. This work is being carried out mainly by W.O'Mullane.

A SIAP web service for the SDSS data was produced. This was requested and tested by R. Williams. T Budavari produced a web service that returns the image information, and

W. O'Mullane calls this in a SIAP compliant service. Many SOAP web services now exist at JHU and these have been listed on a public page. W. O'Mullane used DIME (binary object transfer for SOAP) to transfer images from the SDSS Image Cut-Out Service. A "how to" web page was constructed to show how this may be accessed from Java using the free Axis toolkit.

The Image Cut out service itself is built on SOAP web services and is a good demonstration of this technology. This was shown by M. Nieto-Santisteban at the Pasadena team meeting.

W. O'Mullane and G. Greene (STScI) have been investigating OAI, the Open Archive Initiative. Initial design work has been done to make a Metadata Repository.

T. Budavari , M. Nieto-Santisteban, W. O'Mullane, A. Szalay, and A. Thakar, attended the team meeting in Pasadena.

W. O'Mullane joined the NVO team at JHU in January.

Microsoft Research

No activities to report for this Quarter.

National Optical Astronomy Observatories

In support of WBS 5.2, NOAO convened a meeting of E. Frank (ANL), J. Gray (Microsoft Research), and G. Chisholm (NOAO) to investigate the common elements between High Energy Physics database frameworks (Athena/BABAR/Gaudi) and the SDSS Meta-data/data products/database design. The basic question was how to integrate metadata into the infrastructure. The conclusion was that metadata must be created as a function of the pipeline (i.e., associating the metadata with data products as they are created). The SDSS metadata is inherently built into the database and accessed via SQL queries. Given a generic NVO query language, this model supports a data portal design concept for large synoptic time-domain projects, such as LSST.

In support of WBS 4.1, a proposal was prepared in response to the NSF Middleware Initiative (NMI) entitled "Computational Framework for Astronomical Research (CFAR)" to complement the NVO sponsored work ongoing at NOAO. Specifically, NOAO is building an NVO collaboration to develop generic Grid-based tools to enable the construction of user-defined pipelines for (re-)processing astronomical data. The collaboration includes TeraGrid researchers and principals of the MACHO project to provide both the framework and a high-value science validation dataset. The strategy is to create Grid tools to reprocess the MACHO dataset using new, state-of-the-art algorithms. (Initial experiments with the "image difference" algorithm shows that the detection rate for this phenomenon will increase substantially.) The CFAR proposal is intended to provide funding to develop an infrastructure that extends beyond that described in the initial NVO project, but also to address the immediate NVO needs.

In support of the Science Prototypes (WBS 10.3), M. Fitzpatrick attended the Seattle AAS meeting where he participated in the demonstration of the NVO science prototypes and the VOTOOL: VOTable browser software.

D. De Young participated in NVO Science Demonstration presentations at the AAS meeting, Seattle, 6-9 Jan 2003, and attended the IVOA meeting there on 9-10 Jan. De Young also attended the AVO Science Working Group meeting and AVO Science demonstration, Manchester, UK, 10-22 Jan 2003.

De Young continued working with the US theoretical astrophysics community in order to define a possible theory-based NVO demonstration for the IAU in Sydney. Specific ideas were discussed with P. Teuben (U. Maryland) on using the GRAPE based n-body simulations of globular cluster evolution as a prototype. This is currently under further development.

De Young participated in weekly Executive Committee telecons, biweekly WBS Level 2 telecons, and in the IVOA telecon. He initiated a draft update of the NVO roadmap and circulated a draft press release policy for NVO; both documents are being revised.

National Radio Astronomy Observatory

Most of the effort at NRAO this past quarter has gone into planning development for phase 2 of the NVO data access layer (DAL) and generic client interface. This work included the following activities:

- A call-for-proposals for version 2 of the Simple Image Access interface, due out this summer.
- The conceptual design of a new scalable, distributed, multi-wavelength data analysis framework to link conventional astronomical data analysis and VO, and provide a reference framework for implementing VO data access services. This is documented in a white paper released in draft form in early April.
- Participation in the design of the new registry service, in particular the interface between the registry and data access services, and digital entity naming issues.
- Setting up the IVOA DAL working group. The kick-off meeting of this working group will be at the IVOA interoperability workshop to be held in Cambridge, UK in May.
- D. Tody participated in the NVO team meeting in Pasadena in April, presenting the plans for DAL phase 2.

Work continues on the NRAO archive. The VLA raw data should be available on spinning disk within the next quarter. Arrangements are being made to mirror the entire archive at NCSA. This may serve as a testbed later for data grid technology and scalable pipeline processing and VO analysis.

Discussions were held with representatives of the Australian VO, to coordinate archive development including VO services. A copy of the FIRST survey is being provided to VO-India.

Raytheon Technical Services Company

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

Project-wide. RTSC staff participated in the NVO Project Team meetings and on-line discussion groups, including several IVOA-sponsored e-mail discussion lists. RTSC staff took an action at one of the project team meetings to create and populate a CVS repository for NVO software and demo products. This action was completed; the CVS repository is now available, at <http://nvo.gsfc.nasa.gov/viewcvs/viewcvs.cgi/>, with several entries now added.

WBS 2: Data Models. No significant activity during this reporting period.

WBS 3: Metadata Standards. RTSC staff continued to participate in and support the Metadata Working Group, with particular emphasis on the NVO registries. In addition, staff developed a prototype high-level VOQL (VO Query Language) based on ontology theory. Following numerous discussions on the IVOA-voql web discussion group about infrastructure requirements for this language, a PSL (Problem Statement Language) was created and the XML schema for the PSL was generated. Staff gave a presentation on these developments at the NVO project team meeting. Staff is nearly finished with a white paper describing a federated heterogeneous system for metadata and database query. A prototype top-level Astronomical Data Query Language was designed with XML and Web Services. The draft white paper on this Query Mediator is entitled "Implementing a search across heterogeneous resources using an extensible query layer: application to the Virtual Observatory," which will soon be distributed to the project team for review.

WBS 10: Science Prototypes. RTSC staff is participating in the IVOA online discussions regarding the registry requirements that are needed to support science demos and science use cases. RTSC staff attended the January 2003 AAS meeting in Seattle and answered questions from meeting attendees about the NVO project and science demos. Under the auspices of other research funds, staff is investigating scientific data mining techniques for the NVO, and thereby contributing to those activities and discussions within this project. Staff attended a Data Mining Technologies conference in Feb.2003 and gave a talk on this topic.

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. We expect similar systems to become up to a factor of two cheaper by the end of FY 2003.
- 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. A subset of the catalog has been implemented in DB2, on a 64-processor Sun server. Testing will be done in 2nd quarter 2003.
- Porting of 2MASS and DPOSS collections to Storage Resource Broker version 2.0. The new SRB 2.0 version supports parallel I/O, bulk data registration, direct tape access, and access controls on metadata. These features improve either the performance (data transport rate, data registration rate) or control.
- Initiated discussion with J. Smillie (Australia National University) on the registration of the MACHO image archive into the NVO data grid. Interest in this project has been driven by researchers in the US who want access to MACHO images in bulk.

Smithsonian Astrophysical Observatory

SAO continued to lead the Data Model design (WBS 2.1, 2.2) and the Metadata design (WBS 3.1) efforts. The group at SAO has been working on a requirements and specification document for extending the Simple Image Access Protocol.

Personnel attended or participated in the following meetings:

- January 5-9: AAS Meeting, Seattle
- January 20: AVO demo, Jodrell Bank, UK
- April 3-4: team meeting, Pasadena
- Regular team and metadata telecons

Space Telescope Science Institute

STScI continues to support the public access for NVO standard ConeSearch and SIAP services for GSC-I and II, DSS-I and II, HST observation catalog, and MAST (Multi-mission Archive at Space Telescope) holdings.

Coordination between STScI technical and scientific staff and the JHU SDSS Science Archive team continues for the purpose of technical exchange on NVO technologies currently in development. Biweekly technical exchange meetings continue with presentations on specific topic areas such as Web services and demonstrations of web portal engines such as SkyQuery currently being developed at JHU. Efforts are now being made to establish a local and extended community web-based TWiki for publishing and exchanging information about these projects.

A prototype OAI (Open Archives Initiative) repository was built at STScI to demonstrate technology capabilities that may be used for constructing an NVO registry. Joint telecons were conducted with OAI developer Carl Lagoze to discuss the NVO registry requirements and methods for implementing OAI harvesting along with other metadata group members.

There is ongoing technical exchange with other NVO metadata participants in the first release of a standard XML schema, VOResource, which can be used for defining NVO registry resource and service metadata. The schema was incorporated into the OAI repository for prototyping metadata discovery mechanisms.

For the early prototyping of NVO Registry, STScI and JHU staff are working together to develop a system for loading Registry resources into a SQL Server database and providing web services to support registry data retrieval and discovery. Also in coordination with these efforts, OAI repositories to serve SDSS archive data along with DSS and GSC have been developed and will be integrated into the registry systems. The repositories were built using web services with the hope of further establishing interoperable capabilities.

STScI has continued working with the metadata working group in defining future NVO development in the areas of service metadata specification, registry requirements, and prototyping.

United States Naval Observatory

The USNO Flagstaff Archive server has been upgraded, and now has the capability to return catalogue data in the XML/VOTable format defined by the VO organizations. In addition, a stand-alone ConeSearch code that reads the USNO-B1.0 catalogue was sent to SDSC to help with access to the copy of USNO-B1.0 that was delivered previously. S. Levine attended the March 2003 team meeting in Pasadena.

University of Illinois Urbana-Champaign/ National Center for Supercomputer Applications

R. Plante continues to chair the weekly telecons of the Metadata Working Group. The primary focus of the MWG's agendas this quarter have been on

- Resource registries
- Space-time Coordinates and Regions
- The successor specification for Simple Image Access.

R. Plante and R. Williamson have primarily concentrated on research on resource registries. This has focused on four key fronts:

1. *Resource Metadata Definitions.* R. Williamson created the first version of the Resource and Service Metadata (RSM, Hanisch et al. 2002) in the form of an XML Schema. Based on this, he developed XSL style sheets that convert the RSM into OAI Dublin Core format and the XML Schema into a human-readable metadata dictionary. Plante and Williamson are refining this schema to address the needs of resource registries. Plante, as part of the XML modeling process, has worked at

refining the general RSM data model. He is also prototyping XML Schema authoring styles for producing clear and extensible metadata standards.

2. *Resource Identifier Specification.* Through the MWG, Plante has led the development of a specification for resource identifiers. This included the gathering of requirements and defining the scope. A draft specification is in development.
3. *OAI-PMH Prototyping.* Plante and Williamson have been prototyping a deployable OAI interface that aims at providing a low-cost way for data providers to describe their repositories and associated services. This has included a Web form-based interface that allows curators to enter in high-level descriptions easily.
4. *Registry Prototyping.* Plante and Williamson are participating to the NVO Registry “Tiger Team”, aimed at providing a registry prototype for the Data Inventory Service. We are concentrating on collecting service metadata (via our Web forms) and exposing them via OAI.

In addition, R. Williamson has prototyped some simple Web Services using the Apache Axis package. We expect to do further work in this area in the next quarters when R. Brunner joins our team (with the participation of a new postdoc). Brunner will also be prototyping grid-based data mining for the VO.

Plante continues to refine the Galaxy Morphology Demo in collaboration with E. Deelman (USC/ISI) and J. Annis (Fermilab), preparing it for show at the NCSA All-Hands Meeting in May and Supercomputing '03 in October.

Finally, Plante contributes to the various IVOA working groups. In addition to leading the Metadata Specifications Work Package of the Registry Working Group, he contributes to the working groups for Data Models, UCDS, VO Query Language, and the Data Access Layer.

University of Pennsylvania

U. Penn. staff continued exploration of two designs of new standards for the incorporation of time-series data into a federated database system: (1) FITS extension with NVO compliant metadata (see WBS 3), and (2) insertion into VOTable format. The latter is being implemented on a trial basis using a database containing the MACHO light curve data.

P. Protopapas is preparing a database to host all MACHO light curves using NVO standards. This is an SQL type database that enables light curves to be accessed transparently. It supports SIAP, ConeSearch, and VOTables (for returned results). In addition, Protopapas is preparing a web service that supports the NVO framework, including a capability to provide the union of multiple VOTables.

Penn also participated in the design discussions regarding metadata concepts. P. Protopapas participated in the discussion group for metadata standards and VOQuery language.

University of Southern California (ISI)

Tasks undertaken by USC/ISI during the January 2003-March 2003 quarter:

- Porting Montage to the Chimera/Pegasus framework. ISI ported the Montage application to the Chimera/Pegasus framework. As a result ISI was able to run a small sample Montage computation using a small number of grid 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 and NCSA. ISI is increasing the testbed that was used for the Galaxy morphology computation to include resources from FNAL and NCSA.

University of Wisconsin

No activities to report for this Quarter.

Publications and Presentations

Borne, K. D., "Data Mining with the NVO", invited talk at the Data Mining Technologies conference, Washington, DC (Feb.25, 2003).

Craig, N., Spitz, R., Hawkins, I., & Malina, R., "National Virtual Observatory Outreach: Preliminary Findings of the Qualitative Survey of Artists and Science Museum Professionals," AAS Meeting 201, #53.14

Hanisch, R. J., "The National Virtual Observatory," ESnet Steering Committee, March 2003

Hanisch, R. J., "The NVO Science Prototypes," NASA Headquarters, March 2003, and STScI, February 2003

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

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

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

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

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

Acronyms

| | |
|---------|--|
| AAS | American Astronomical Society |
| ADEC | Astrophysics Data Centers Executive Committee (NASA) |
| 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 |
| DAGMan | Directed Acyclic Graph Manager (Condor) |
| DAL | Data Access Layer |
| DAML | DARPA Agent Markup Language |
| DARPA | Defense Advanced Research Projects Agency |
| DM | Data Model |
| DOE | Department of Energy |
| DPOSS | Digitized Palomar Observatory Sky Survey |
| DTD | Document Type Description |
| EDG | European Data Grid |
| 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 |
| GSC | Guide Star Catalog |
| HEASARC | High Energy Astrophysics Science Archive Center |
| HTM | Hierarchical Triangular Mesh |
| 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 |
| MACHO | MAssive Compact Halo Object |
| MAST | Multi-mission 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 Archive Initiative |
| OASIS | On-line Archive Science Information Services (IRSA) |
| OGSA | Open Grid Services Architecture |
| OIL | Ontology Inference Layer |
| PB | petabyte |
| PSL | Problem Statement Language |
| Q | quarter |
| QSO | Quasi-Stellar Object |
| RC | Replica Catalog |
| RDF | Resource Description Framework |
| RLS | Replica Location Service |
| ROME | Request Object Management Environment |
| RSM | Resource and Service Metadata |
| RTSC | Raytheon Technical Services Corporation |
| 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 |
| STScI | Space Telescope Science Institute |
| SWG | Science Working Group |
| TB | terabyte |
| UCD | Uniform 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 System 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 |