RUSSIAN ACADEMY OF SCIENCES

Institute of Informatics Problems
IPI RAN

Special Astrophysical Observatory
SAO RAN

Information Infrastructure of the Russian Virtual Observatory (RVO)

Briukhov D.O., Kalinichenko L.A., Zakharov V.N.
(IPI RAN)

Panchuk V.E., Vitkovsky V.V., Zhelenkova O.P.
(SAO RAN)

Moscow 2005

Briukhov D.O., Kalinichenko L.A., Zakharov V.N., Panchuk V.E., Vitkovsky V.V., Zhelenkova O.P. **Information Infrastructure of the Russian Virtual Observatory (RVO)**. M.:IPI RAN, 2005. – 152 p. – ISBN 5-902030-09-9.

This publication is the final report of the project supported by the Russian Foundation for Basic Research, grant № 04-07-90083, dedicated to the analysis of the Russian Virtual Observatory information infrastructure (RVOII) as one of the first steps in achieving the objectives of RVO. RVOII is aimed at representation of information in various problem domains of astronomy to support scientific research.

The report also contains: analysis of different kinds of astronomical information resources; analysis of the modeling, technological and architectural recommendations of the IVOA; analysis of correspondence and sufficiency of the IVOA standards for the identified RVO activities; analysis of existing components that can be re-used for RVOII implementation. Based on the analysis performed, the information architecture of RVO has been defined.

For specialists in astronomy and information technologies.

Эта публикация представляет аванпроект информационной инфраструктуры Российской виртуальной обсерватории (РВОИИ), разработанный при поддержке Российского Фонда Фундаментальных Исследований, грант № 04-07-90083. РВОИИ ориентирована на представление информации в различных проблемных областях астрономии для поддержки научных исследований.

Аванпроект содержит также: анализ разнообразных видов астрономических информационных ресурсов; анализ технологических и архитектурных рекомендаций IVOA; анализ соответствия и достаточности стандартов IVOA для различных видов деятельности РВО; анализ существующих компонентов, которые можно использовать при реализации РВОИИ. На основании проведенного анализа, разработана информационная архитектура РВО.

Для специалистов в областях астрономии и информационных технологий.

FOR FURTHER INFORMATION PLEASE CONTACT:

Institute of Informatics Problems RAS

44-2 Vavilov str., Moscow, 119333, Russian Federation

Tel. 7 095 3324933

E-mail: leonidk@synth.ipi.ac.ru

ISBN 5-902030-09-9

Table of contents

1 Introduction

This publication is the final report of the project supported by the Russian Foundation for Basic Research (RFBR), grant № 04-07-90083, dedicated to the analysis of the Russian Virtual Observatory (RVO) information infrastructure as one of the first steps in achieving the objectives of RVO. The RVO Information Infrastructure (RVOII) project started in Spring 2004 and ended in December 2004 as a joint effort of the Special Astrophysical Observatory of RAS (SAO RAS) and the Institute of Informatics Problems of RAS (IPI RAS). RVOII is aimed at representation of information in various problem domains of astronomy to support scientific research. Brief characterization of the final report content follows.

The report starts with a list of RVO objectives that were taken into account for the RVOII project. Basic concepts of representation of problem domains in natural sciences are discussed. A brief analysis of different kinds of astronomical information resources, including surveys and catalogs in optic, infrared and radio spectral ranges, archives of observations, centers of astronomical data, as well as robotic telescopes, is provided. In this analysis the astronomical resources of Russia are given separately. The report gives significant attention to the analysis of the state and decisions of the advanced Virtual Observatory (VO) projects (NVO, AstroGrid, Euro-VO). Analysis of classes of astronomical research problems that should be supported by RVOII constitutes an important part of the report.

The report provides a detailed analysis of the modeling, technological and architectural recommendations of the International Virtual Observatory Alliance (IVOA). Complementary to the IVOA architecture and standards, for RVOII new components are proposed: subject mediators that should provide scientists with facilities for problem definition and solving over multiple heterogeneous information resources, digital libraries for astronomical science education, robotic telescopes as specific information resources. The report shows that introduction of the subject mediator concept can provide for more consistent treatment of the IVOA data modeling, data access level and other activities. An approach for the subject mediator architecture is defined, an example of a simple mediator for distant galaxy discovery is given.

Based on the analysis performed, the information architecture of RVO has been defined. Various kinds of the RVO activities to be supported by RVOII have been identified. An analysis of correspondence and sufficiency of the IVOA standards for the identified RVO activities has been undertaken. Significant number of activities has been identified that are not supported by any standard. Another kind of analysis consisted in identification of existing components that can be re-used for RVOII implementation. Around 50 free components have been identified and recommended for RVOII.

Basic RVOII principles have been formulated. The RVO Information Infrastructure has been specified. The conceptual infrastructure for the initial RVO prototype with two data centers has been proposed. Structure of the detailed work packages for the RVOII development according to the five directions identified has been specified. The work program presupposes creation of working groups for each of the directions identified.

Strategically the program of the RVOII development assumes tight coordination of work with International VO activities. In particular, the work program presupposes active participation of the RVOII developers in the IVOA Working Groups.

2 Objectives of the RVO project

Nowadays there exist a lot of large digital archives of observations related to specific instruments used by various astronomical observatories in the world. Among them are the archive of the Hubble Space Telescope, the archive of the Chandra X-Ray Observatory, the Sloan Digital Sky Survey (SDSS), the Two Micron All Sky Survey (2MASS), the Digitized Palomar All Sky Survey and many others. Each of these archives is of great value, but for research (e.g., requiring studying of objects in several spectral ranges, or analyzing the objects variability) the data obtained by different instruments are required. The volume of digital sky survey in one spectral band is of several terabytes. It is unlikely that every scientist studying such problems could support a copy of the required data. Many new instruments are under development for studying of our Galaxy and extragalactic objects. All the data obtained would contain the unprecedented quantity of information for the studying of the Universe evolution provided that these data could be processed in an integrated way as if they were collected in one data store. Various archives containing observations of tens of millions of astronomical objects as the results of spectral or monitoring surveys are accessible through the Web. Taken together, they provide on the orders of magnitude more information than can be obtained from a single instrument. Besides that, practically most of the astronomical literature is available through the Internet establishing also relationships between the publications and observations.

Celestial objects radiate energy in all bands of the electromagnetic spectrum, including radio, infrared, optical, ultraviolet, X- and Gamma-Rays. Observations in each of these bands provide important information on the nature of the objects observed. One and the same object may manifest itself completely different in different spectral bands. To solve the problems of integrated usage of the astronomical data, the astronomical community is developing a new approach to work with the observation data based on the concept of the Virtual Observatory (VO). According to this approach, various IT instruments are under development for the integrated access to heterogeneous distributed archives and catalogs of data as well as to the

computational facilities located close to the data locations to minimize data transfer overhead. Real telescopes will perform new surveys filling up existing repositories, or monitoring of interesting events or objects for the more detailed analysis. Astronomers will get an opportunity to discover unknown phenomena getting access to vast spectral and temporal data.

Virtual observatories are under intensive development by various astronomical institutions in the world (among them are the National Virtual Observatory (NVO) in USA, the Astrophysical Virtual Observatory (AVO) in the European Union, the Canadian Virtual Observatory (CVO), the Japanese Virtual Observatory (JVO) and many others). Joint efforts are coordinated by the International Virtual Observatory Alliance (IVOA) as well as by the Commissions of the International Astronomical Union (IAU). The Russian Federation also announced the work on the Russian Virtual Observatory and joined IVOA.

Main objectives of the RVO project (http://www.inasan.rssi.ru/eng/rvo/, [VRVOI]) have been defined as follows:

- to provide the Russian astronomical community with the facilities of integration of the Russian astronomical resources into the VO;
- to provide the Russian astronomical community with the facilities of integrated access to the data accumulated in the International astronomical data resources and in the Russian resources constituting together the tangible digital representation of the Universe in various spectral bands (in the opposite way, to provide the International astronomical community with an access to data accumulated in Russia (or probably even in the FSU countries));
- to provide the Russian astronomical community with the facilities of problem domains definition for solving of various classes of the astronomical problems, computational facilities, facilities for information analysis and data mining, facilities for automation of scientific research in astronomy;
- to develop and support a set of standards agreed with the international community and providing for the interoperability of heterogeneous data and facilities listed above for the problem solving;
- to develop strategically important classes of astronomical problems based on the VO technology and develop processes and mediators for the respective research support;
- to develop organizational measures for development and usage of the VO technology in Russia agreed with the international community, for coordinating of the Astronomical Data Centers in Russia and abroad, for coordination of research based on the VO technology;

- to develop a set of measures for creation of RVO as an important educational resource for the Russian Universities;
- to fill in the recently formed gap in the achieved level of development and use of the VO technology in Russia and in the rest of the World;
- to form in Russia the sustainable community of astronomers actively using VO in their scientific research;
- to contribute for the high level of research based on VO technology in Russia in the strategically important areas of astronomy.

The RVOII project report dedicated to the analysis of the RVO information infrastructure is one of the first steps to achieve the objectives listed. The report contains analysis of the kinds of activities of the astronomical community that should be supported by RVO, short survey of various kind of astronomical information resources, overview of the state and organization of the advanced VO projects in the world, analysis of the state of the International standards for VO and their sufficiency for RVO, analysis of approaches to the organization of problem solving in astronomy on the basis of the subject mediation technique. Based on the analyses listed, the RVO information infrastructure is proposed as well as the structure of the work packages to implement this infrastructure.

3 Representation of problem domains in natural sciences in information systems

The gradual evolution of information systems from currently dominated frameworks in science to more knowledge-based organization is expected. Information in the information systems for science and education (acting as *collective memories*[1] (CM) converging sources of various kinds) should be organized differently than in accordance with the conventional database or digital library metaphor [KSMDL]. Papers, journals, books, textbooks, courses are not good information entities any longer. Scientists have spent centuries to reach well-defined structures, concepts and theories in various science domains. These definitions are more suitable as a guiding principle for information structuring and search in collective memories. Knowledge-based collective memory in a domain of natural science should include domain terminology and

[1] The Technical Committee on Digital Libraries of the Institute of Electrical and Electronics Engineers Computer Society (TCDL of IEEE-CS) to define what digital library is, uses general term "(digital) *collective memory*" to emphasize the convergence of sources of various kinds. Collective memory development faces challenges in several areas, including storage, classification, and indexing; user interfaces; information retrieval; content delivery; presentation, administration; preservation, etc.

concept definitions, material system descriptions, definitions of various theories and models, observable (measurable) characteristics of real world objects, description of methods and instruments for observation, measurement, observation and simulation data, data analysis results, problem definitions and methods of solution, algorithms and programs, simulators. Integration of such information is driven by scientific and educational needs (Fig. 1). The range of information required includes also scientific researches reported both formally in journals and informally in Web sites in the domain, curricula and courseware materials, lectures, access to remote scientific instruments, tools, the results of educational research, raw data for student activities, as well as related multimedia (image, audio, or video) banks. National and international CM development in various science domains now are actively underway applying integration of data bases, digital libraries, data grids, and persistent archives.

Results of large-scale computer simulations and other theoretical predictions are to be published in CM in a form compatible with real observations. Many archives, both observational and theoretical, already make their *information products* available over the Internet [LTHEO]. To *publish an information product* means to make it available in an archive through services that are accessible via a CM supplied Internet portal. *Discovery and manipulation* are basic operations in CM over publications. More specific examples of service kinds include query and browsing services, analysis services (e.g., data mining, inference), simulators.

Requirements for scientific results publishing:

- To allow independent checks of conclusions based on theoretical results;
- To allow further analysis by third parties based on the published data;
- To allow comparisons with similar results/methodologies or with the corresponding data by observers/theoreticians;
- To make theoretical results more easily accessible and understandable for observers;
- Journals may allow/require links to actual information products and/or software used in published work;
- To allow querying of publications, real and simulated information products in a uniform manner (joint queries on a structured content items and on metadata – on observations and publications);
- To allow invariants for observable classes, treating them as interpretations of theories (models), triggers watching for inconsistencies of observations and theoretical models;
- Referees may insist they be able to reproduce certain results through CM services.

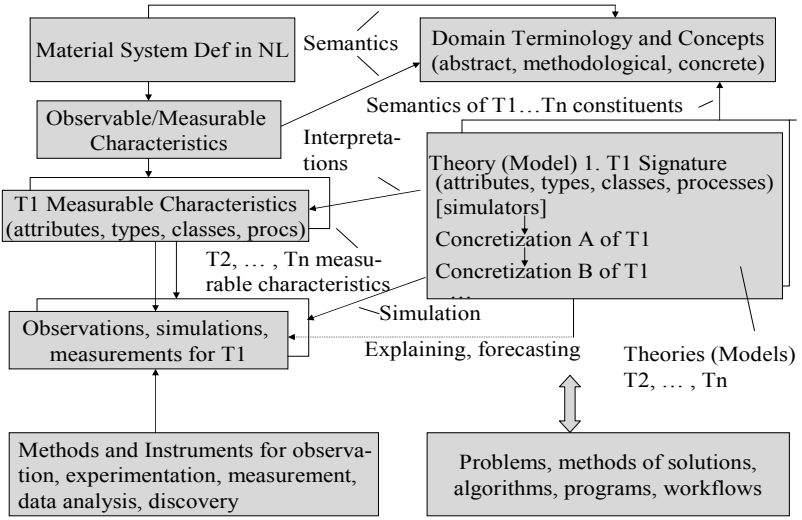

Figure 1. Subject domain definition in natural science

Federation of theoretical with observational archives and a uniform (canonical) model of the *metainformation* describing the information products are required to enable the comparison of observational with relevant theoretical information products. The Federation is enabled through the cross-matching of catalogs, where objects are identified as the same. Large-scale cross-matches will require substantial quality control and human insight in order to produce usable results.

The publication is meaningful when it is associated with a *curation* mechanism to assert quality. It is difficult or impossible to discover by what process some piece of data got into the CM. Scientific data are typically «curated» – they are derived from other data, but considerable effort in annotation and correction goes into their construction. Thus these data have added value and are in turn used as sources for data for other projects. There may be several curation stages from the source data. While curated data are extremely valuable, understanding the original sources for a data element in a curated database is a problem. Data should carry with it an indication of how it was obtained. Data *provenance* is closely related to a number of other issues that have already been identified as central to information systems: *data survivability, data cleansing, data quality.*

Wide *community* around the CM is to be formed to provide for the CM development, governance, collecting and processing information, standards development. The community includes researchers, members of professional societies, information providers, representatives of industries, educators,

students. CM *sustainable development and gradual evolution* is a serious issue. Large CM should function as non-profit organizations. Building sustainable and scalable communities where e-science and education best practices and information are shared (national and global communities, domain (discourse)-oriented communities) requires specific attention and organizational measures. CM subdisciplines include storage, curation, archiving, classification and indexing, administration, cataloguing, registering, preservation, mediation, information discovery, content delivery, portal services, presentation, annotation, evaluation, personalization, recommendation, and copyright management.

Complex scientific problems increasingly require collaboration between teams of researchers distributed across space and time applying different methods, models, vocabularies, and philosophies. Effective collaboration remains dependent on researcher ability to consolidate uniform subject domain specifications, co-create meaningful explanations, and on being able to describe the knowledge that supports these explanations.

Systematic investigations are still required to analyze efficiency of research innovations including group work on research problems, customization of information for different roles of researchers, educators, learners, interaction of scientists having different roles with the required subject domain, domain-based (curriculum-based) access to information objects in CM, work with re-use of existing information objects for different kinds of job, etc.

Several important issues require separate discussion, e.g.:

- sustainability and economic issues are crucial for the CM development;
- preserving national language/theoretical-school/historical identity in research vs globalization of CM (global CM vs national CMs that somehow interoperate, content in different natural languages, interdomain semantic work).

4 Information Resources in Astronomy

4.1 World-wide resources overview

Information resources used by astronomical community can be classified as follows: sky surveys and catalogs; observatory and space mission archives, digitized plate collections; data centers. Up to the middle of the 20th century, main astronomical resources included photographic plate collections and printed catalogs. Technological progress with launching of space missions and appearance of new generation telescopes, observation range extension, rise of detector sensibility and coordinate measurement precision just increased acquisition rate of digital data. As a result, observation digital archives of ground-based and space observatories, archive data centers with web access to

data appeared. In the 1990's next generation digital surveys making by ground-based telescopes brought Tbytes of data to the community. Total astronomical data volume is estimated as a few hundred of Tbytes. Digital data amount doubles every 1.5 years. Large volumes and distribution of the resources pushed an issue of the VO as an information infrastructure based on web-services and distributed computing. The projects of further synoptic surveys started in the beginning of the 21st century. The surveys (PB volumes) are directed to the temporal domain research (from minutes to years). Few samples of world-wide resources will be overviewed in this section.

4.1.1 Optical surveys and catalogs

First surveys and catalogs

The first surveys of the entire sky were photographic ones. The First Palomar Sky Survey (POSSI) [APALS] was carried out in the 1950's using the 48-inch Oschin Schmidt telescope at the Mount Palomar Observatory. During the 1970's, the U.K. Schmidt telescope, nearly identical to the Oschin telescope, carried out the Southern Sky Survey. In the early 1980s, Palomar's Oschin telescope was upgraded and carried out a second sky survey called POSS II [DPALI]. The surveys gave rise to a few digital sky surveys – the Digitized Sky Survey 1 (DSS1, http://archive.stsci.edu/dss/acknowledging.html), the Second Digitized Sky Survey (DSS2, http://archive.eso.org/dss/eso-dss.html), the SuperCosmos Sky Survey (SSS) [RCOSS] and to catalogs such as the Guide Star Catalog (GSC) [RSPCE], the United State Naval Observatory catalog (USNO) [USNO], the Astronomical Precise Machine catalog (APM, http://www.ast.cam.ac.uk/~mike/apmcat/overview.html).

In order to support the Hubble Space Telescope (HST) operations and provide a service to the astronomical community, the Space Telescope Science Institute (STScI) has digitized the photographic plates and created the DSS**.** The DSS comprises a set of all-sky photographic surveys in five bands. The 6.5-degree x 6.5-degree plates were scanned using a microdensitometer.

Digital Survey	Volume (TB)	Step (mk)	Pixel size (arcsec)	Plate size (GB)	Plate pixel size
DSS1	0.6	25	1.7	0.4	14000
DSS2	3	15	1	1.1	23040
SSS	10	10	0.7	2	

DSS2 consists of the higher resolution scans of several more plate collections in the red, blue, visible and near infrared bands. SSS has about the same features but a pixel step is smaller than DSS size.

The most frequent use of the digital sky surveys are to generate finding charts prior to, or during observing runs, search for optical counterparts of objects discovered at other wavelengths, the resulting statistical multi-

wavelength studies, searches for rare or previously unknown types of objects (objects with unusual colors and/or morphological structure). From an individual object research astronomers got an ability to apply statistical methods. Currently the surveys are one of the most utilizable resources. The catalogs and surveys are provided for on-line access and use in NVO and VO-compliant projects. For example, SkyView (http://skyview.gsfc.nasa.gov/cgi-bin/titlepage.pl) is a resource on the net generating images of any part of the sky at wavelengths in all bands from radio to gamma-ray.

Sloan Digital Sky Survey

The next information and technological level achieved in digital survey carrying out is the Sloan Digital Sky Survey (SDSS, http://www.sdss.org). To map more distant, fainter objects, then the DSS a telescope with lenses 2.5 meters (100 inches) across was built for the SDSS. The survey maps one-quarter of the entire sky, determining the positions and absolute brightnesses of hundreds of millions of celestial objects. It measures the distances to more than a million galaxies and quasars, performs a redshift survey of galaxies, quasars and stars, provides images, imaging catalogs, spectra, and redshifts for download.

SDSS data releases:

Rel.	Date	Size (TB)	Images (M)	Spectra (K)
DR3	27.09.2004	3	141	478
DR2	15.03.2004	2	88	330
DR1	15.06.2003	1	53	186
ERD	06.06.2001	0.2	14	54

DR3 data are distributed via the Catalog Archive Server (CAS), an SQL database with fast search capabilities, and the Data Archive Server (DAS), a collection of FITS images and tables containing the outputs of the imaging and spectroscopic reduction pipelines. The total SDSS volume is 44TB. Innovative IT decisions were required for SDSS data processing and dataflow, realization of access to images and catalogs. For example, SkyQuery (http://www.skyquery.net) and SkyServer (http://www.sdss.org) created a prototype of a federated database application, implemented using a set of interoperating web services. This is a good demo of how typical applications in the Virtual Observatory might look like.

Large Synoptic Surveys

The next two optical surveys mark a new age in the Universe research. These surveys will add a bulk variability object research on the different temporal concordance from minutes to a few years to the expanding of observational spectral ranges. The advent of synoptic surveys presents extreme

opportunities; as an illustration, consider the SDSS, which over the course of 5 years represents a factor of a million increase in information over previous surveys; however, the LSST will amass a SDSS every 3 nights. The exploration of the temporal domain results in data sets that are not just more voluminous than before, but far richer and more complex. This presents challenges to all aspects of astronomy: data gathering, distribution, reduction, analysis, storage, archiving, dissemination and mining.

The **Palomar-Quest (PQ) Variability Survey** [MPALQ] is a new, digital, synoptic sky survey, using large area CCD camera at the Palomar 48" Oschin Schmidt telescope (2003). A major new feature of the PQ survey will be many repeated observations of the same portions of the sky, enabling researchers to find not only objects that move (like asteroids or comets), but also objects that vary in brightness, such as the supernova explosions, variable stars, quasars, or cosmic gamma-ray bursts. Time baselines is ranging from minutes to days, months, years (repeat scans) to decades (cross-match to DPOSS, SDSS, etc.).

The existing PQ pipeline is geared to complete processing of a night's data by the next day. The data rate is 2.45MB/s and with a monthly average of 10 nights' observing, PQ produces ~ 1TB of data/month. The survey is a science and technology precursor/testbed for the LSST and other major synoptic sky surveys in the future.

To illustrate how PQ will make use of VO technologies in an integrated fashion, consider one of the pipeline systems Palomar Real Time Transients Discovery System. It will produce real time (within four minutes of the data being taken) alerts of transient events (e.g. supernovae). The system data flow has a few steps. The PQ nightly catalogs will be compared with older catalogs from the survey itself, as well as other surveys and archives, using NVO infrastructure and methodology. Positions of possible transients will be cross-matched with those of known variable sources, known asteroids, etc. The goal is to provide detections of potentially exciting sources in real-time, via email alerts and a dedicated website. A key challenge will be to deal with this abundance of data in an effective manner – maintaining a high completeness in terms of the interesting variable and transient sources discovered and doing it in real time. A number of advanced statistical and Machine Learning techniques will be explored to this end.

The next example of case study with PQ is the quasar discovery workflow that will use the P48, P60, P200, Keck telescopes and discovery system for quasar detection.

The **Large Synoptic Survey Telescope (LSST)** [LSSTO] will be a large, wide-field ground-based telescope designed to deeply image of the entire visible sky every few nights, reach 24 V mag in <15 sec over ~10 sq deg field with time resolution: 20+ sec, limiting magnitude: 26.5 AB magnitude. LSST will generate 100 Petabytes of data (~2 Terabyte per hour). And the survey data rates are:

- Over 3 GBytes /sec peak raw data from camera;
- Real-time processing and transient detection: < 10 sec;
- 0.6 GB/sec average in pipeline;
- Real reduction requires ~ 100 Tflops peak;
- Data rate is comparable to ATLAS on LHC.

LSST science goals: dark matter/dark energy via weak lensing and supernovae; Galactic structure encompassing local group; dense astrometry over 20000 sq.deg: rare moving objects; gamma ray bursts and transients to high redshift; variable stars/galaxies: black hole accretion; 5-band 27 mag photometric survey: unprecedented volume; Solar system probes: Earth-crossing asteroids, comets, TNOs.

4.1.2 Infrared and radio range surveys

Two Micron All Sky Survey

The Two Micron All Sky Survey (2MASS, http://www.ipac.caltech.edu/2mass) carried out a uniform digital imaging survey of the entire sky in three near infrared bands. The primary survey data products were derived from the approximately 24.5TB of raw digital images. Data were released to the astronomical community: a digital Image Atlas with 4.1 million calibrated FITS images; a Point Source Catalog (PSC) containing accurate positions and photometry for ~471 million sources; and an Extended Source Catalog (XSC) that contains accurate positions, photometry and basic shape information for ~1.6 million resolved sources.

2MASS Atlas Images are accessible via the 2MASS Image Services that are administered by the Infrared Science Archive (IRSA, http://irsa.ipac.caltech.edu). The scientific value of 2MASS is documented in over 600 publications. Data rate of 2MASS data processing system is 35 GB/night that required highly efficient automated software “pipeline” that converted night’s raw imaging data into calibrated images and extracted source lists.

Faint Images of the Radio Sky at Twenty-cm

The Faint Images of the Radio Sky at Twenty-cm (FIRST, http://sundog.stsci.edu) [WFIRS] is a survey to produce the radio equivalent of the POSS over at 1365 and 1435 MHz. The survey will result in a catalog of discrete sources as well as 65000 images. The astrometric reference frame of the maps is accurate to 0.05". The survey area has been chosen to coincide with that of the SDSS; at the m(v)~24 limit of SDSS, ~50% of the optical counterparts to FIRST sources will be detected. The images are 7.1 Mbytes to cover a 0.45 square degree area of the sky. A FIRST catalog for the north and

south Galactic caps (811,117 sources), was derived from the 1993 through 2002 observations.

NRAO VLA Sky Survey

The NRAO VLA Sky Survey (NVSS, http://www.cv.nrao.edu/nvss) is a 1.4 GHz continuum survey covering the entire sky north of -40 deg declination. The principal NVSS data products are images and a catalog. A set of 2326 continuum image cubes each covering 4 deg x 4 deg with three planes containing the Stokes I, Q, and U images. The rms uncertainties in right ascension and declination vary from < 1 arcsec for relatively strong (S > 15 mJy) point sources to 7 arcsec for the faintest (S = 2.3 mJy) detectable sources. The catalog of discrete sources on these images is over 1.8 million sources in the entire survey.

4.1.3 Archives of observations

Archives of observations include observatory and space mission archives as well as digitized plate collections. They contain raw data of observations obtained with different instruments. Data structures in archives are highly heterogeneous.

The European Southern Observatory Archive

The European Southern Observatory (ESO). The ESO/ST-ECF Science Archive Facility (http://archive.eso.org). The Archive currently holds about 21 TB of active HST and ESO data. The ESO Archive Database contains all observations performed with the ESO telescopes. Currently available in the archive are the data from the VLT, NTT, MPG/ESO 2.2m, and the ESO 3.6m telescopes. Images and spectra are stored in the archive in an unreduced (raw) form together with the calibration and auxiliary data. The archive facility offers access to data collected by a number of survey projects carried out within or in collaboration with ESO or the STScI.

The ESO and the Space Telescope-European Coordinating Facility (in collaboration with the Canadian Astronomy Data Centre, CADC, http://cadcwww.dao.nrc.ca) have implemented a number of innovative features for the ESO/ST-ECF Science Archive Facility that have since become part of a set of 'minimum requirements' for modern astronomical archive systems. These features include previews, web interfaces to query the database, on-the-fly re-calibration, association of exposures, data mining.

The National Radio Astronomy Observatory

All astronomical data from the National Radio Astronomy Observatory (NRAO, http://www.nrao.edu/) radio telescopes will be archived and cataloged in this unified system, with the data accessible by direct ftp downloads. The archive currently contains raw data and catalog tables with the meta-data from

each observation (position, time on source, frequency, bandwidth, etc). Currently the data archive contains approximately 6 TBytes of data products from the telescopes (VLA, VLBA and GBT). There is about 10 TBytes of raw VLBA data in the existing tape archive that are being copied onto the hard disk archive.

Hubble Space Telescope Archive

The Hubble Space Telescope (HST) Data Archive (HDA, http://archive.stsci.edu) contains science data from all completed HST observations and calibration files. Its volume is over 20 Terabytes of data. In addition to all the science data sent to observers and all the calibration reference files, the archive contains engineering files that may be useful for diagnosing some questions about observations.

Spitzer Space Telescope

The Spitzer Space Telescope (formerly SIRTF, the Space Infrared Telescope Facility is exploited since 25 August 2003, http://www.spitzer.caltech.edu). During its 2.5-year mission, Spitzer will obtain images and spectra by detecting the infrared energy, or heat, radiated by objects in space between wavelengths of 3 and 180 microns. MIPS and IRAC (SIRTF devices) both generate about 20 GB/day of data.

The SIRTF Wide-Area InfraRed Extragalactic Legacy survey (SWIRE) will survey ~65 square degrees of high latitude sky in all 7 SIRTF imaging bands tracing the evolution of dusty star-forming galaxies evolved systems and AGN in the context of cosmic structure formation from z~3 to the current epoch. Key Scientific Goals: the evolution of both actively star-forming and passively-evolving galaxies to determine the history of galaxy formation and the evolution of their clustering in the key redshift range from $0.5 < z < 3$ over which much of cosmic evolution has occurred. The survey will be dominated by more than 10^5 luminous infrared galaxies; up to 40,000 with $z > 2$; 1 million early-type galaxies (up to 400,000 with $z > 2$); 30,000 classical AGN, and as many as 250,000 dust-obscured QSO/AGN.

International Ultraviolet Explorer

International Ultraviolet Explorer (http://www.vilspa.esa.es/iue/iue.html) was a joint project between NASA, ESA and PPARC (formerly SERC). IUE was the longest and most productive astronomical space Observatory, the first general user UV space Observatory. IUE was into orbit on January 26 1978 and was turned off on 30 September 1996. 18.7 years of IUE operations returned 104470 high- and low-resolution images of 9600 astronomical sources from all classes of celestial objects in the 1150-3350 A UV band. 3585 publications have been published in the refereed scientific literature using IUE results.

The images obtained with IUE were collected and world-wide accessible through National Hosts with ESA's INES (http://ines.vilspa.esa.es/) system. The IUE Data Archive remains the most heavily used astronomical archive, each IUE spectrum has already been used six times. During IUE's life, more than 1000 European observing programs were conducted from Villafranca, returning more than 30000 spectra from about 9000 targets, extending from Comets to far away quasars at the early days of the Universe and covering a brightness range of 10 orders of magnitude (extending from M_V=-4 to M_V=21). The INES is mirrored into 21 countries including Russia.

Chandra X-ray Observatory

NASA's Chandra X-ray Observatory (Chandra, http://chandra.harvard.edu), which was launched and deployed by Space Shuttle Columbia on July 23, 1999, is the most sophisticated X-ray observatory built to date. Chandra is designed to observe X-rays from high-energy regions of the universe, such as the remnants of exploded stars. Chandra's angular resolution and celestial location capability is ideal for corresponding optical survey, to allow optical identification of the majority of the X-ray sources.

4.1.4 Data centers

Data centers intending for collecting, storage and distribution of astronomical catalogs provide for the users different services (storage, data access and more advanced services binding with on-line data analysis) and embed innovative IT. As example of large centers is Centre deDonnées astronomiques de Strasbourg (CDS), the Canadian Astronomy Data Centre (CADC), the Astronomical Data Analysis Center (ADAC, Japan, http://dbc.nao.ac.jp). There is a net of national data centers. Archive centers like Multimission Archive at Space Telescope (MAST, http://cdsweb.u-strasbg.fr), the Infrared Processing and Analysis Center (IPAC, http://www.ipac.caltech.edu) have the same functions and additional data processing. The centers contain large and popular catalogs and surveys, archives, famous astronomical site mirrors, library databases. Content volume in CADC is about 55TB, in ADAC – 11TB. More details on certain data centers follow.

CDS

The Centre de Donnees de Strasbourg (CDS, http://cdsweb.u-strasbg.fr) has been providing database services to the astronomical community for more than 25 years. The CDS provides access to 4000 catalogs with help of VizieR [OVIZI], Aladin [BALAD], SIMBAD [ESIMB] information systems. Also, the CDS provides the images of 2MASS and DSS-I, DSS-II, HST, SSS, FIRST, NVSS, Merlin, XMM-Newton, Chandra, all images are provided by SkyView (HEASARC, http://heasarc.gsfc.nasa.gov).

Vizier is a database containing about 4000 catalogs (10 millions rows in relational database). It is mirrored into 8 other data centers. VizieR is a database grouping in an homogeneous way thousands of astronomical catalogs gathered for decades by the CDS and participating institutes. Several query interfaces are currently available, making use of the ASU protocol, for browsing purposes or for use by other data processing systems such as visualization tools.

The Aladin interactive sky atlas is a service providing simultaneous access to digitized images of the sky, astronomical catalogues, and databases. The driving motivation is to facilitate direct, visual comparison of observational data at any wavelength with images of the optical sky, and with reference catalogues. The system processes ~70000 database queries per month from ~6500 nodes.

The SIMBAD astronomical database provides basic data, cross-identifications and bibliography for astronomical objects outside the solar system. SIMBAD can be queried by object name, coordinates, other criteria (filters), and lists of objects. SIMBAD contains 3,334,996 objects, 8,727,968 identifiers, 151,102 bibliographical references, 4,611,718 citations of objects in papers.

MAST

MAST (Multimission Arkhive at Space Telescope) is the primary archive and distribution center for HST data, distributing science, calibration, and engineering data to HST users and the astronomical community at large. Over 100000 observations of more than 20000 targets are available for retrieval from the Archive. The one contains: >40 TBytes (20 TB HST); ingest rate: 15 GB/day (5.3 TB/yr); retrievals: 52 GB/day (19 TB/yr);14 years of HST data; plus GALEX, FUSE, IUE, DSS, GSC2, VLA FIRST.

IRSA

The NASA Infrared Science Archive (IRSA) at IPAC curates data from six NASA infrared missions, including the 2MASS 10 TB full resolution image data set and the 1 billion source working catalog. The highly-scaleable architecture and algorithms developed at IRSA support production of these massive data sets, their curation in the archive, their accessibility by the astronomical community, and their interoperability with remote data sets. IRSA archives and serves the 2MASS science products: IRSA User Services; Custom query engines for individual data sets; General search engines; Gator – catalog search engine; OASIS – Java data fusion; Atlas – multiple data sets; RADAR – scaleable inventory service; Web interface – returns html page; http request – returns FITS file, or subset of source returns catalog; NVO compliant request – returns VO compliant XML structure.

4.1.5 Surveys and Robotic Telescopes

A robotic telescope is an automatically operating astronomical instrument which carries out celestial object observations. An automation capabilities of instruments vary largely, but usually include: automatic switch-on and switch-off of the instrument at a given time; analysis of weather conditions; targeting, search and tracking of an object; tuning of telescope devices and setting of detector modes; registration, acquisition and transfer of digital data; diagnostics and logging of instrument work; user notification about alarm and in-service situations.

Robotic telescopes are used for observations in different spectral ranges for science researches and education. Examples of robotic telescopes applications include deep extragalactic survey, follow up photometry of gamma bursts, searching of micro gravitational lenses, spectroscopy of our Galaxy stars.

More than 100 robotic telescopes exist around the world, about 50 new projects are being developed. Mirror diameters of robotic telescopes vary from 25sm to 2.5m. For example, the On-line Observatory Project (Japan) includes three small telescopes 20, 25, 45 sm diameters placed in Tokyo and Osaka. The system is used for education. Students can prepare observational programs in distant mode and get results at convenient time.

The Bradford robotic telescope (46sm diameter, Tenerife, Spain) is used mainly for education. Tens of English and Spanish schools have access to it. The telescope is also used for following up of optical bursts.

The Liverpool automatic telescope. The telescope is fully computer automated and for its maintenance does not need permanent technical staff. An observation process is completely automatized beginning from dome opening, a fast object guiding, changing of any one of 5 devices, can work at strong wind. There is a project to install similar six telescopes around the world for continuous observations.

British astronomers created a global net of robotic telescopes RoboNet. One of the goal of the system is searching like Earth planets. The net consists from three telescopes placed in UK, Australia and Canaries. The program system developed at the Liverpool John Moores University and high-speed links enable to control them from a single center. A geographical location of the telescopes allows following up from objects without intervals. The system has on the fly respond on transient events and carries out sky surveys with automatic data processing and object detection.

The eSTAR Project (http://www.estar.org.uk/) attempts intelligent agent technologies to carry out resource discovery, submit observation requests and analyze the reduced data returned from a network of robotic telescopes. The agents are capable of data mining and cross-correlation tasks using online catalogues and databases and, if necessary, requesting additional data and

follow-up observations from the telescopes on the network. The network consists of a number of autonomous telescopes, and associated rapid data reduction pipelines, connected together using Globus middleware.

There are two fundamental ideas behind eSTAR which make it a unique project. The first is to treat both telescopes and databases in as similar a fashion as possible, both being made available as a resource on the "Observational Grid". The second is that the main user of that grid should not be humans making observing requests, but should be intelligent agents. The design is analogous to computational grids, with no overall supervisor, giving the system scalability with multiple agents talking to multiple nodes.

The agents were deployed in the field onto a non-robotic research class telescope, UKIRT. Observation requests are made by the user's intelligent agent to an agent embedded at the Joint Astronomy Centre (JAC), where the request in Robotic Telescope Markup Language (RTML, http://alpha.uni-sw.gwdg.de/~hessman/RTML/) was automatically translated to the JAC's internal Telescope Observation Markup Language (TOML, http://omp.jach.hawaii.edu/schema/TOML) format.

All aspects of an observation program at the JAC are either software readable or software controllable via the Observation Management Project (OMP, http://omp.jach.hawaii.edu/). This allows the embedded agent to fully specify an observation, which is placed in the queue as a high priority target of opportunity which is seen when the observer next requests an observation. When the data is taken by the observer it is automatically reduced in real time by the ORAC-DR system which returns the fully reduced data to the embedded agent, which forwards the result back to the user's intelligent agent. To the user's agent it is irrelevant that there is a human in the loop.

It is intended to broaden the abilities of the eSTAR network by deploying the agents onto more telescopes, including the Liverpool and Faulkes Telescopes, and to distribute an agent tool kit to allow the easy construction of intelligent agent by astronomers. This is essential for further progress and widespread adoption of the technologies that have been developed. It is important that, as much as possible, federated databases and telescopes share a common interface.

4.2 Russian astronomical resources

A list of the Russian astronomical resources provided by various Russian organizations is given at http://www.inasan.rssi.ru/rus/rvo/rus_res.html.

4.2.1 Resources of the Special Astrophysical Observatory of RAS (SAO RAS)

CATS

The CATS database (http://cats.sao.ru/) [VCATS] is a support system of astrophysical catalogs. The informational system was created as RATAN-600 observation run support system. The system is widely used by astronomical community (2500-3000 requests per month from Russian and foreign users). The system contains more than 260 catalogs with total volume of 1GB, number of records is more than 5.8 millions. It carries out selection, context searching, cross-matching, creation of source spectra from multi frequency catalogs.

SAO RAS Archive

The archive **(**http://www.sao.ru/oasis/cgi-bin/fetchru**)** [ZARCH] contains raw and calibration data obtained with the 6-m, Zeiss-1000, Zeiss-6000 optical telescopes and RATAN-600 radio continuum data. The archive current state is in the table bellow. The archive contains 16 local archives obtained by different optical devices. Service functions include web interface for inquiries by date and device, quick-look file headers and FITS images, download by ftp.

SAO RAS observation archive (CD-disks):

Archive	CD	Data rate per date	Volume	Files
Optical	111+111 copies	60.4MB	65.4GB	127100
Radio	7	4.3MB	3.6GB	45921
User's	30 copies			

The Archive of Spectral, Photometric and Interferometric Data (ASPID)

The observational data archive (http://alcor.sao.ru/db/aspid/) contains results of observations conducted by the laboratory team on the 6-m Telescope. It includes observational data obtained with Integral Field Spectrograph (MPFS), Multi Object Fiber Spectrograph (MOFS), 'Long Slit Spectrograph' (LS), Fabry-Perrot Interferometer and BTA Optical Reducer (SCORPIO). Observational data presented in ASPID were obtained on the 6-m Telescope for different observational programs and are included into data archive as coming from Telescope in FITS format. The author of the observational program and the identification of the program can be found in each detailed data set description presented in database. Data volume is about 60GB.

The Information system “Evolution of radio galaxies” and ‘Big Trio’ archive

The project of the informational system creation on the problem of evolution of radio galaxies, as a part of the "Big Trio" project aimed on studying of distant radio galaxies, has been developed. This system (http://sed.sao.ru/) [VRGAL] allows a user to operate with simulated curves of spectral energy distributions (SED) to estimate ages and redshifts by photometry data (42GB).

The solar radio observation archive

The archive contains daily RATAN-600 observations of the Sun. Gif pictures and FITS files since 1997 are located on the site http://www.sao.ru/~sun.

CD/DVD disk collection

CD/DVD disk collection contains digital surveys and catalogs (DSS1, USNO-A2, 2MASS, 122 disks, data volume is about 80GB) and astronomical software systems (6).

4.2.2 Resources of INASAN

The Centre of Astronomical Data of the Institute of Astronomy RAS (CAD)

The center contains astronomical catalogs regularly obtained from CDS since 1980. CAD (http://www.inasan.rssi.ru/rus/cad/index.html) has a catalog fund on 240 CD disks containing astronomical data and catalogs (228), publications (2), astronomical software (5), others (5). The center mirrors important astronomical databases (VizieR, ADS, INES). CAD carries out preparing, checking and transferring catalogs and data tables from articles published in Russian astronomical journals. The staff does data analysis of large catalogs and creates service software for them.

4.2.3 Resources of the Pulkovo Main Astronomical observatory of RAS

System of astrometric database

The system of astrometric database includes three databases containing results of processing of photographical and CCD observations obtained with 26-inch refractor and normal astrograph.

Databases with solar activity observations

Extended time series of Solar Activity Indices (ESAI) are stored in a database (http://www.gao.spb.ru/database/esai/) including observational, synthetic and simulated sets to study Solar magnetic field variations and their influence on the Earth. ESAI includes monthly sunspot areas (1821-1989), yearly sunspot areas (1821-1994) and yearly mean latitudes of sunspots (1854-1985);

Pulkovo database of sunspot magnetic fields

Pulkovo database of sunspot magnetic fields (http://www.gao.spb.ru/database/mfbase/gindex.html) contains long-term daily observations. There is a simple web interface for data requests.

Combined database of sunspot magnetic field

The database (200MB) contains data of magnetic field observations obtained in seven observatories.

Bulletin 'Solar data'

Bulletin 'Solar data' (http://www.gao.spb.ru/russian/index.html) contains solar activity data obtained by observations of 6 observatories and solar radio emission from 7 observatories and sunspot groups since 1996.

Photographical library

Photographical library (http://www.gao.spb.ru/russian/index.html) is a collections of astronomical plates obtained with the Pulkovo telescopes since 1893. About 40000 plates were scanned with a digitized machine. Approximate data volume is about 3TB.

4.2.4 Resources of SAI

HyperLeda mirror

HyperLeda (http://www.sai.msu.su/hypercat/) is an information system for astronomy. It consists of a database and tools to process the data according to the user's requirements. The scientific goal consists on studying of physics and evolution of galaxies. The database contains about 3 million objects; out of them 1 million are certainly galaxies. Since 2003, HyperLeda is distributed as part of MIGALE. HyperLeda is now developed in the general frame of the Virtual Observatory.

General Catalogue of Variable Stars

This GCVS edition (http://www.sai.msu.su/groups/cluster/gcvs/gcvs/) contains data for 37470 individual variable objects, the catalogue of variable

stars in external galaxies contains 10979 variable objects, the catalogue of extragalactic supernovae includes 984 objects. Catalogues are available from Sternberg Institute via anonymous ftp, where they are stored as zip and text files (ftp://ftp.sai.msu.su/pub/groups/cluster/gcvs/gcvs).

Sternberg Astronomical Institute Supernova Catalogue

The current version of SAI Supernova Catalogue (http://www.sai.msu.su/sn/) presents data on 2872 extragalactic supernovae discovered since 1885 until June 15, 2004 and on their parent galaxies.

Catalogue of Interacting Galaxies by Vorontsov-Velyaminov

(http://www.sai.msu.su/sn/vv/)

Photographical library

Photographical library containing 46000 wide-field plates obtained with different telescopes. The plates will be scanned in collaboration with CAD.

4.2.5 Russian Robotic Telescopes

Among the Russian robotic telescopes [VRAST] are a) the project MASTER (http://www.pereplet.ru/lipunov/202.html#202) – Mobile Astronomical System of Telescope Robots. The instrument is working since 2002. It consists of three optical cameras. The prototype is fully automated and connected with Internet. Main telescope applications include optical transient observations, sky surveys (up to 19^{th} magnitude) and asteroid searching; b) the Automated Astronomical Complex of the Novosibirsk State University (AAC, http://vega.inp.nsk.su/index.php3?introduction). Any amateur can have access to it after registration. The AAC site provides convenient navigation facilities for observations which include descriptions of observation methods and astronomical software for data processing and analysis. AAC is used for student training, starry heaven excursions for amateurs and for astrophysical researches. AAC is equipped with a reflector with 315 mm mirror and a solar telescope with 100mm objective.

5 Analysis of Projects of Virtual Observatories

Several astronomical institutes such as the National Virtual Observatory (NVO, US), the European Astrophysical Observatory (AVO, Euro-VO), the Canadian Virtual Observatory (CVO) carry out the works on VO. On the East the VO projects are being developed in Australia (http://www.aus-vo.org), China (http://china-vo.org/en/index.php), India (http://vo.iucaa.ernet.in), Japan (http://jvo.nao.ac.jp/index-e.html). Joint efforts are coordinated by the International Virtual Observatory Alliance (IVOA) [IVOAL] and the IAU Commissions.

In this section a brief overview of the projects considered to be the major players in the area (NVO, Euro-VO and AstroGrid) will be given. The objectives of the overview is to provide the intensions of different projects, project participants and financing, standards applied, information infrastructure used, state of the art achieved.

5.1 NVO

In August 2001, the US National Science Foundation awarded five-year funding to a collaboration "Framework for the National Virtual Observatory", under its Information Technology Research program [NVOBS]. NVO's objective is to enable new science by greatly enhancing access to data and computing resources. The 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. The NVO will establish this environment through the use of high-performance computing, large-scale databases, 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. The NVO is developing tools that make it easy to locate, retrieve, and analyze astronomical data from archives and catalogs worldwide, and to compare theoretical models and simulations with observations. Though initial objectives of the project was a research about how Virtual Observatories could be made, now NVO provides quite advanced facilities for astronomers as well as contributes to the IVOA standards. NVO team members both lead and participate 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-led the development of an IVOA standards process, a process recently endorsed by the IVOA Executive Committee.

More than 70 people at 20 organizations and institutions are participating in the NVO project. These organizations include: Caltech, Canadian Astronomy Data Centre/Canadian Virtual Observatory, Carnegie-Mellon University/University of Pittsburgh (CMU/UPitt), Fermi National Accelerator Laboratory (FNAL), High Energy Astrophysics Science Archive Research Center (HEASARC), Johns Hopkins University, Microsoft Research, National Optical Astronomy Observatories (NOAO), National Radio Astronomy Observatory (NRAO), Raytheon/ADC, San Diego Supercomputer Center, Smithsonian Astrophysical Observatory, Space Telescope Science Institute, United States Naval Observatory, University of Illinois-Urbana/Champaign/National Center for Supercomputer, Applications (UIUC/NCSA), University of Pennsylvania, University of Southern California

(USC/ISI), University of Wisconsin. Involvement of various organizations into the project can be found in the NVO September 2003 Annual Report [BFNVO]. Main scientific, technological and educational objectives of the NVO can be summarized as follows (the order in the list reflects time ordering for the duration of the project):

1. Scientific objectives
    - Incorporate requirements from the theoretical astrophysics community, developing a science demonstration based on theoretical simulations of globular clusters;
    - Collect and integrate science requirements for access to spectral data. Expand data access capabilities with additional Cone Search and SIAP services;
    - Make direct comparisons of observed and simulated data, with focus on Globular clusters[2];
    - Execute multiple large science runs that analyze of the contents of entire image archives;
    - 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;
    - Search for outliers and clusters in large federated datasets via data mining algorithms;
    - Find faint, variable, and very extended objects in large federated datasets via data mining in the image domain;
    - Compare large (TB-scale) theoretical simulations with observational data;
    - NVO-enabled science should be visible in the peer-reviewed literature;

[2] This demonstration compared a library of N-body simulations of the evolution of globular clusters with observational data. Direct comparisons of color-magnitude diagrams from both observations and calculations were possible, thus permitting the determination of which theoretical models best reproduced the evolution of a given cluster.

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

2. Technological objectives
   - Define VOTable standard 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;
   - 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 scalability of algorithms and data access methods. Incorporate support for virtual data products;
   - 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;

- 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;

3. Educational objectives
   - Incorporate EPO (Education and Public Outreach) 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;
   - Promote interactions with the EPO community through the deployment of an NVO portal for outreach;
   - Encourage and aid resource managers in converting data with EPO potential into EPO ready data;
   - Seek interaction with NSF NSDL initiative for development of curricula based upon NVO resources.

In accordance with the objectives listed, the activities and accomplishments in the NVO project are impressive. The following issues have been investigated and prototyped:

1. System Architecture: relationships to Grid components, computational facilities (processing, bulk data storage, network access, security, authentication) as well as Digital library integration;
2. Registries (resource metadata, publishing and harvesting protocols, query protocols), replication, synchronization, maintenance, revision control, and curation;
3. Data models (High-level (image, spectrum, time series, event lists, visibilities, catalogs, simulations, data quality), Low-level (measurement, quantity, uncertainty, relationship), Descriptors and ontologies (UCDs), Space-Time and regions);

4. Data Access Layer (Data access services (catalog, image, spectrum, time series, visibilities, etc.), Data representation (VOTable, etc.), Data provider/consumer implementations and end-to-end testing);
5. Query Language (Low-level: Astronomical Data Query Language, Mid-level: VOQL and OpenSkyQuery/OpenSkyNode, High-level: Complex queries);
6. Web and Grid Services (Web Services (SOAP, WSDL, etc.), Grid Services (OGSA), Computational resource management, Virtual data, Application and service integration with Grid);
7. Applications (Data location services, Cross-correlation services, Image combination, registration, Visualization tools and services, Theory, Statistical analysis, Data mining, Outlier identification, Interfaces to/from legacy software systems).

In particular, for computational facilities, to support installation of the NVO testbed on the TeraGrid, replication of NVO selected surveys onto TeraGrid storage systems, execution of Atlasmaker to create standard image projections for surveys, execution of the Montage mosaicing software, fitting of quasar spectra, development of n-point correlation functions of galaxies, and development of a Cosmic Microwave background grid.

5.2 EURO-VO

The European Virtual Observatory (EURO-VO) project [EURVO] is planned as an integrated and coordinated program of work designed to provide the European astronomical community with the data access, research tools and systems, research support, data interoperability standards, data-flow practices and data centre coordination, necessary to enable the exploration of the digital, multi-wavelength universe resident in European and international astronomical and astrophysical data archives. EURO-VO is planned as a four year project (starting probably in 2005) that should result in complete set of standards, tools and organizational infrastructure constituting VO for Europe. EURO-VO will be built on the results such projects as AVO [ASTVO], AstroGrid and others and will be compliant with the IVOA standards. The EURO-VO program seeks to support and deploy Virtual Observatory (VO) capabilities to data centers and observatories across the entire electromagnetic spectrum. It will therefore be closely coupled with the two other major integrating and networking activities for astronomy in FP6: OPTICON and RADIONET. EURO-VO will act as a natural hub for coordination and integration of the new, GRID-enabled, VO research infrastructure that will be essential to the success of future large European community programs in astronomy (e.g. ALMA, OWL, SKA and Planck).

The following organizations form the Euro-VO consortium:

- French VO, as represented by the Centre de Données astronomiques de Strasbourg (CDS), Strasbourg, France;
- European Southern Observatory, Garching, Germany;
- European Space Agency, Paris, France;
- UK AstroGrid Consortium, as represented by the University of Edinburgh, UK;
- German Astrophysical Virtual Observatory (GAVO), as represented by the Max Planck Institute for Extraterrestrial Physics (MPE), Garching, Germany;
- Istituto Nazionale di Astrofisica, Rome, Italy;
- Nederlandse Onderzoekschool voor Astronomie, Leiden, The Netherlands;
- Laboratorio de Astrofísica Espacial y Física Fundamental, Madrid, Spain.

The EURO-VO will consist of three new organizational structures [EUTEC, EUNET, EUINT] which will meet the objectives of the total work program and which will provide a platform for a long term European VO research infrastructure and capability. There are three fundamental EURO-VO objectives:

- EURO-VO-Objective 1: Technology take-up and full VO compliant data and resource provision by astronomical data centers in Europe;
- EURO-VO-Objective 2: Support to the scientific community to utilize the new VO infrastructure through dissemination, project support, tool prototyping and VO facility-wide resources and services;
- EURO-VO-Objective 3: Further development and refinement of VO technologies to meet new scientific challenges.

The three EURO-VO structures that will meet these objectives are:

- The EURO-VO Data Centre Alliance (DCA): a collaborative and operational network of European data centers which, by the uptake of new VO technologies and standards will publish data, metadata and services to the EURO-VO and which will provide a research infrastructure through the adoption and application of GRID-enabled processing and storage facilities;
- The EURO-VO Facility Centre (VOFC): an organization that provides the EURO-VO with a centralized registry for resources, standards and certification mechanisms as well as community support for VO technology take-up, VO dissemination and scientific program support using VO technologies and resources;

- The EURO-VO Technology Centre (VOTC): a distributed organization that coordinates a set of research and development projects on the advancement of VO technology, systems and tools in response to scientific and community program needs.

The first step for the VO projects worldwide is to develop the standardized framework. Once in place, the framework must be taken up by data providers to allow them to publish their holdings and make their services and facilities available. The framework will then empower scientists and developers to provide new tools for research and to undertake new research programs to tackle complex astronomical and astrophysical problems. There are, therefore, three fundamental tasks to be undertaken to make the VO a successful research infrastructure:

1. The completion and deployment of a VO technical infrastructure;
2. The uptake of this infrastructure by data providers;
3. The support of the research and development community to utilize this infrastructure and data content to discover new knowledge and build new capabilities;

Data centers can intervene in different critical aspects of the VO, such as:

- to provide astronomers with easy, long term access to observation archives;
- to develop expert data centers which provide services to the community;
- to improve efficiency by sharing expertise and reusing experience, techniques and tools when applicable;
- to improve data quality by providing progressively calibrated data;
- to develop and implement visualization and data analysis tools;
- to implement simulation data in the VO, i.e. with the possibility to re-use them and to compare them to observational data.

The DCA Network main objective is to create a network of European data centers sharing basic interoperability functions (as defined incrementally by the VOTC Network and JRA 1). This has two aspects:

- Promotion of common protocols and standards, to spread good 'VO practice' among European data managers, which will allow for publication of data with full VO interoperability. This will be achieved: (1) by organizing one DCA Network Workshop per year ('Interoperability meeting'), open to members of partner organizations and to data centre managers from other European countries and from candidate and associated countries; (2) by providing technical support to data centers, in collaboration with VOFC.

- Implementation of basic interoperability functions in data centers supervised by the national nodes, to enable a first level of networking at European level. Taking advantage of the DCA Committee membership (national representatives), the program of action, defined by looking for 'European added-value', will be well phased with national VO priorities. The DCA Committee will meet regularly to assess the status of VO interoperability standards and tools and to discuss the list of actions to be undertaken in the different data centers overseen by the national nodes. The DCA Committee will also review the status of the networking implementation. It will if necessary propose to include new partners.

Basic development principles:

1. To assess new technologies and study the feasibility of their incorporation in Euro-VO;
2. To create designs of new infrastructure components based on those new technologies;
3. To create designs of science user tools and data mining services;
4. To develop trial versions of new infrastructure components, tools, and data mining services and to test them;
5. To decide what new interoperability standards are required, and to define those standards with international partners;
6. To liaise with the larger Euro-VO structure, gaining refreshed versions of science functionality and architecture, and feeding back component test results, designs, and trial components for demonstration suites;
7. To liaise with computer science, IT industry, and related applications projects in order to mesh with larger standards and to save work wherever possible.

At the end of the Design Study, the following deliverables are planned:

- a series of study reports in the areas of grid services, the semantic web, ontology, data mining, visualization, agents, workflow, and distributed storage;
- a final architecture design for Euro-VO;
- a series of design documents for selected tools;
- a series of design documents for new infrastructure components;
- a series of internationally agreed astronomical interoperability standards in all necessary areas (ontology, workflow, etc);
- trial implementations of new infrastructure components;

- interface specifications, to allow external projects to use components, data centers to publish data, and for and external user development of new tools and services;
- a study assessing the financial implications of Euro-VO in terms of construction and operations;
- design document for the technical operation of user support and training for the Euro-VO;
- Euro-VO project plan.

Significant interest and uptake from several classes of users of the resulting infrastructure is expected: (1) The very large community of astronomical end-users: the Euro-VO infrastructure, the datasets that populate it, and the tools available through it, will become an integral part of the daily life of almost every astronomer; (2) The community of professional data centers. They will need to structure their archives in a Euro-VO compliant manner, install the necessary components, and construct and publish data services; (3) The data creation facilities, every new telescope, every new instrument is to be built with Euro-VO in mind. This includes consortia of astronomers who construct major new scientific data sets using those facilities; (4) Science tools writers. Some of this activity will go on inside data centers, some in specialized groups developing new data mining algorithms and so on, and some will be undertaken by interested individuals. Tools will also include major theory tools – simulations, photo-ionization codes, etc.; (5) The educational system and the general public – easy access to the best data and tools at students' desktops.

Licensing Policy: The Euro-VO, in common with other VObs initiatives, is working in the public domain, publishing their products into the public domain, using open-source principles. Reports resulting from this design study will be published by the Euro-VO. Software elements produced during the design study will be released under the IVOA Public License (which is currently under development).

The astronomy GRID network is at present based on a large set of heterogeneous, distributed, observatory archives, compilation databases and electronic journals. A first set of Interoperability standards are available to allow information networking and integration, and more are being developed. But tools and results of theoretical astronomy are in general not yet accessible: some code is available on demand or on line, with some documentation, some simulation results are obtainable, but there is no framework allowing publication and usage of codes and results in a standardized manner. This would be a major step forward, allowing scientists to re-use the tools for new simulations, to compare the results of different models, and to compare and visualize simulation and modeling results to observational data available in the astronomical GRID network. Electronic links already exist between services

inside this information GRID: in particular, scientists can navigate among bibliographic resources, from the ADS bibliographic database, to electronic journals and compilation databases, and vice versa. Observatory archives have begun to record the journal articles that cite their data. This very useful information is at present gathered by librarians or scientists checking all published papers for citations of the observations made at their observatory.

Euro-VO project emphasizes the concept of Data Centers. DCA introduces additional level in the information organization hierarchy. Consequences of such decision for RVO require serious analysis when more information on DCA will be available. The technological state of the art preceding Euro-VO is shown by AVO 2005 Demo (January 2005) that is considered as a step towards the Euro-VO (http://www.euro-vo.org/twiki/bin/view/Avo/AvoDemo2005). IVOA minimal standards needed for AVO 2005 demo include:

- Registry Interface V0.9;
- UCD 1+;
- ADQL V0.7.4;
- SIA V1.0;
- SSA V0.9;
- SkyNode (standard WSDL);
- VOTable V1.1.

VOTech part of Euro-VO [EUTEC] is funded now for 5.5M US$ for 2005-2007 considering AstroGrid, ESO, French and Italian groups.

5.3 AstroGrid

The AstroGrid project [AGRID] aims at producing a working datagrid for key selected databases, with associated data mining facilities, by late 2004. It is part of the world-wide drive towards the concept of a Virtual Observatory (VO), and can be seen as the UK contribution to this vision. The aim of the UK AstroGrid project is to focus on short term deliverables, both relevant application tools and the federation, by the data centers that manage them, of key sky survey datasets, namely: (1) SuperCOSMOS, Sloan, INT-WFC, UKIRT WFCAM, XMM-Newton, Chandra, MERLIN and related VLA datasets; (2) SOHO and Yohkoh; and (3) Cluster and EISCAT. The differences between the data types involved in these federations means that each brings distinct challenges, whose solutions will shed light on generic problems to be faced by the developing global "Virtual Observatory". A three year program, started in 2001 and estimated to cost £4M (on the assumption of existing funding to establish the archives and provide on-line storage) which will add value to existing UK astronomy resources in the short term, as well as positioning the community strongly with respect to wider Grid initiatives, in astronomy and beyond.

The goals of the AstroGrid project are :

- A working datagrid for key UK databases;
- High throughput data mining facilities for interrogating those databases;
- A uniform archive query and data-mining software interface;
- The ability to browse simultaneously multiple datasets;
- A set of tools for integrated on-line analysis of extracted data;
- A set of tools for on-line database analysis and exploration;
- A facility for users to upload code to run their own algorithms on the data mining machines;
- An exploration of techniques for open-ended resource discovery.

The AstroGrid in various ways is both wider and more focused than other initiatives. It is wider in that it covers astronomy, solar physics, and space plasma (solar terrestrial) physics, and covers all wavelengths from radio to X-ray. The project is also part of a coherent UK e-science program, with links to projects in particle physics, bio-informatics, and basic grid technology development.

AstroGrid is however focused in that it aims to develop something recognizably like a working VO on a short timescale, so that science can start getting done and technological lessons can be learned. This requires concentrating on selected datasets. The project priority is to develop a virtual observatory capability to support efficient and effective exploitation of key astronomical data sets of importance to the UK community: for example data from WFCAM, VISTA, XMM-SSC, e-MERLIN, SOHO and Cluster. It seems clear that good data curation, archive management, and datamining services all need to be closely linked together. AstroGrid is therefore a partnership formed by UK archive centers and astronomical computer scientists.

Typical workflow to handle a fairly simple query according to AstroGrid looks as follows (http://www.star.le.ac.uk/%7Ecgp/ag/dbstatus.htm):

1. The User specifies the URL of a convenient Data Portal, and uses its query interface to specify the results or processing required. The interface may be a simple forms interface, but (as noted later) we are likely to devise an astronomically-oriented query language to make it easier to specify the more advanced queries;
2. The Data Portal analyses the query and sends it to the nearest VO Resource Registry. It is expected that nearly all queries will result in a request for information from the Registry, but those which are very specific and name the data archive to use can perhaps be handled directly by the portal;

3. The Resource Registry then replies with a list of the archive sites which may have relevant information: in order to avoid missing information this list will always err on the side of generosity;
4. The Data Portal then interrogates each of these remote archives to find the protocols they can handle (using WSDL) and subsequently seeks information on the detailed contents of their datasets (table descriptions, column UCDs and other metadata);
5. The Portal then formulates appropriate queries and send them encapsulated in SOAP where possible, but for compatibility with archives using non-VO interfaces, it may be necessary to handle queries using ASU or raw CGI form parameters for some time;
6. These results will, if all the archives conform to VO standards for interfaces, be returned in some standard form, for example for small tabular datasets the VOTable format. But the results from different sites may have different formats, in which case the VO Data Portal should be able to assist by carrying out transformations or data reformatting steps to make the results easier for user to absorb.

This workflow (defined in 2002) probably is quite early understanding of query processing in AstroGrid. Reliable information on current understanding of the issue is not available.

By the end of 2004 AstroGrid planned to have constructed a fully functional prototype VO system, linking together major data centers in the UK. Information concerning the Beta test program for AstroGrid is available at http://wiki.astrogrid.org/bin/view/Astrogrid/BetaTesting.

The following components (see Fig. 2) have been included into an early Beta test program:

- Portal: developed with Cocoon, and including portal pages for Workflow, MySpace Explorer, Data Viewer, with a simple interface to the AstroPass component;
- Registry & AstroMQ: implemented with IVOA schema standards, with a registry management portal page;
- Data Centre, Dataset Access, Job Control: allows a query of, and data return from a dataset/database;
- MySpace: enabling the creation of files in MySpace (not tables), a MySpace registry and data mover component.

AstroGrid is to be considered as a part of Euro-VO. Funding in 2001 – 2004: 8.9 M US$, in 2005 – 2007 (Astrogrid 2): 9.6M US$. Staff: 24 FTE. Planned components for AstroGrid 2 are shown on the Fig. 2.

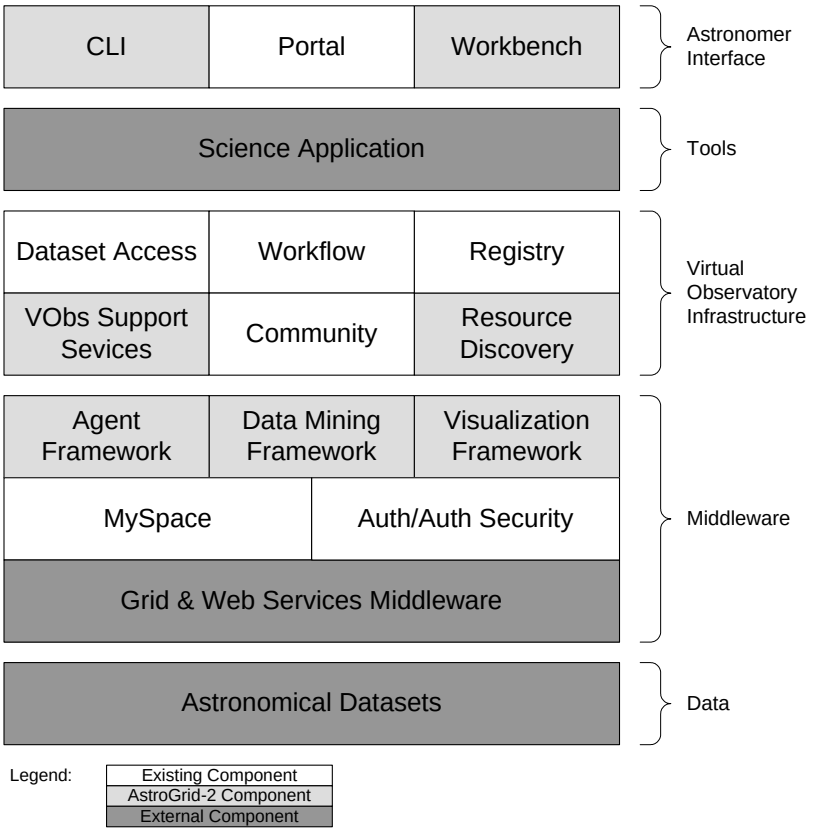

Figure 2. AstroGrid existing and planned components

5.4 IVOA

Various VO projects are funded through national and international programs, and all projects work together under the International Virtual Observatory Alliance to share expertise and develop common standards and infrastructures for data exchange and interoperability. The Astrogrid, AVO and NVO projects took the opportunity to formally announce the IVOA [IVOAL] in 2002.

International Virtual Observatory Alliance Partners:

- AstroGrid (UK) (http://www.astrogrid.org);
- Australian Virtual Observatory (http://avo.atnf.csiro.au);
- Astrophysical Virtual Observatory (EU) (http://www.euro-vo.org);
- Virtual Observatory of China (http://www.china-vo.org);

- Canadian Virtual Observatory (http://services.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/cvo/);
- German Astrophysical Virtual Observatory (http://www.g-vo.org/);
- Hungarian Virtual Observatory (http://hvo.elte.hu/en/);
- Italian Data Grid for Astronomical Research (http://wwwas.oat.ts.astro.it/idgar/IDGAR-home.htm);
- Japanese Virtual Observatory (http://jvo.nao.ac.jp/);
- Korean Virtual Observatory (http://kvo.kao.re.kr/);
- National Virtual Observatory (USA) (http://us-vo.org/);
- Russian Virtual Observatory (http://www.inasan.rssi.ru/eng/rvo/);
- Spanish Virtual Observatory (http://laeff.esa.es/svo/);
- Virtual Observatory of India (http://vo.iucaa.ernet.in/~voi/).

Many of the IVOA projects have active Science Working Groups (SWG) consisting of astronomers from a broad cross-section of the community representing optical, radio, high energy, space and ground-based astronomy. In some cases, IVOA projects have cross-membership of these groups. The common focus of SWGs is to form a clear picture of the scientific requirements for an operational virtual observatory. These requirements are a mix of new technologies and algorithmic capabilities as well as new standards that address fundamental issues of publishing data in the IVO (e.g., guidelines for describing all the aspects of data quality). Individual SWGs have identified the need for a design reference mission for the IVO which will capture the set of tools astronomers will need to do new science in the IVO as well as defining initial science cases and projects that can be run in the IVO to test and refine capabilities.

6 Information Infrastructure Forming Standards

Some of the emerging information infrastructure forming standards are mentioned in this section as quite important for the RVO infrastructure.

6.1 OAI Protocol for Metadata Harvesting

The OAI-Protocol [LOAIF] for Metadata Harvesting (OAI-PMH) defines a mechanism for harvesting records containing metadata from repositories. The OAI-PMH gives a simple technical option for data providers to make their metadata available to services, based on the open standards HTTP (Hypertext Transport Protocol) and XML (Extensible Markup Language). The metadata that is harvested may be in any format that is agreed by a community (or by any discrete set of data and service providers), although unqualified Dublin Core is

specified to provide a basic level of interoperability. Thus, metadata from many sources can be gathered together in one database, and services can be provided based on this centrally harvested, or "aggregated" data. The link between this metadata and the related content is not defined by the OAI protocol. It is important to realise that OAI-PMH does not provide a search across this data, it simply makes it possible to bring the data together in one place. In order to provide services, the harvesting approach must be combined with other mechanisms. For VO optional SOAP version to augment HTTP Get standard is also known (http://nvo.stsci.edu/voregistry/STOAI.asmx).

NVO supports now several metadata registries:

- STScI/JHU Registry (http://nvo.stsci.edu/voregistry/);
- NCSA Registration Portal (http://nvo.ncsa.uiuc.edu/nvoregistration.html);
- Caltech Carnivore (http://mercury.cacr.caltech.edu:8080/carnivore/).

NVO Registries contain over 3000 records.

For registering for VO, minimal information required should contain organization data with ID and collection data sufficient for users to access it via a Browser. Additionally, browser-based services, traditional CGI services, Web Services can be provided on registering. On the next level, one or more VO standard services (such as Cone Search, Simple Image Access, SkyNode) can be implemented

Examples of collections registered in NVO are:

- NCSA Astronomy Digital Image Library (http://adil.ncsa.uiuc.edu/);
- Spectra: Spectrum Services for the VO (http://voservices.net/spectrum/).

6.2 Web Services

A Web service is defined as a standardized way of integrating Web-based applications using the XML, SOAP, WSDL, and UDDI open standards over an Internet protocol backbone. XML is used to tag the data, SOAP is used to transfer the data, WSDL is used for describing the services available, and UDDI is used for listing what services are available. According to W3C, a Web service is a software application identified by a URI, whose interfaces and bindings are capable of being defined, described, and discovered as XML artifacts. A Web service supports direct interactions with other software agents using XML-based messages exchanged via Internet-based protocols.

Web services are no different from other middleware services, with the exception that it should be possible to invoke them across the Web and across organizations. With Web services, designers and developers are led to think in the direction that “everything is a service”. At the same time, Web services do

not need to be accessed through the Internet. It is perfectly possible to make Web services available to clients residing on a local LAN.

Service description in conventional middleware is based on interfaces and interface definition languages. The semantics of the different operations, the order in which they should be invoked, and other (possibly non-functional) properties of the services are assumed to be known in advance by the programmer developing the clients. In Web services such implicit context is missing. Therefore, service descriptions must be richer and more detailed, covering aspects beyond the mere service interface. XML is used as a common meta-language to describe different aspects of services. Such aspects include: interface definitions (applying WSDL), Web service conversation language (e.g., BPEL), service properties and semantics (UDDI). UDDI specification defines standard APIs for publishing and discovering information into service directories.

From the point of view of Web services, the communication network is hidden behind a transport protocol. Web services consider the use of a wide range of transport protocols, the most common one being HTTP. A standard way to format and package the information to be exchanged is provided by the Simple Object Access Protocol (SOAP). A Web service can be implemented by invoking other Web services, possibly provided by different organizations.

Web service technology is one of the basic ones for the VO architecture. VO services are defined as Web services.

6.3 Grid

Grid technology is a natural way of designing the IT infrastructure for e-Science [VODGR]. e-Science refers to science that is enabled by the routine use of distributed computing resources by end-user scientists. It is most effective when is applied to distributed global collaborations involving large numbers of people and large-scale resources. The following classes of Grids might be helpful for this report [RVODG]:

Compute/File Grid

This model is of great importance to support utility computing or computing-on-demand. For example, this model is needed by particle physics (LHC) where the data is many petabytes per year of individual events that need to be analyzed independently and then looked at collectively to find signals of new science or to measure cross-sections. The Globus team calls this a data grid and so to avoid confusion with database-centric applications in e-Science, it is given here the compute/file grid label. The functional capabilities needed by compute/file grids are well understood and illustrated by the work of the European Data Grid (EDG). Characteristic of this style of Grid are resource brokering both within collections of computers and the meta-schedulers and planners between separately managed computer subsystems. Some variant of

distributed file and storage (tape) systems are needed as is the ability to create and manage data replication (caching)

Information Grid

This type of Grid involves integration of large scale distributed data repositories. Information Grids are typified by applications like the virtual observatory and bioinformatics where the typical service is accessing a database. These grids start with a service-centric view as this approach is already popularized by the well developed web interface to databases. OGSA-DAI is a highlight of the UK e-Science program and is the key Grid enabling technology in this area (will be considered later in more details). Information Grids require basic registry and information services with rich metadata required to annotate both the data and the different resources.

Hybrid Grid

Hybrid Grid is a combination of Information and Compute/File Grid emphasizing integration of experimental data and simulations. This hybrid Grid links Information and Compute/File Grids and can be expected to be of growing importance, specifically in the VO area where simulations gain large attention.

Semantic Grids

It is expected as Grids become increasingly endowed with metadata they will naturally evolve into Semantic Grids. In a service-oriented architecture for the Grid, metadata is associated with services. For example, a service might contain metadata that describes its capabilities, interfaces, provenance, performance, security and access policies, and so on. This metadata could be used, for example, by a workflow service to decide how two services might interact, or by a resource broker to decide how to schedule a service-based application. Services can also publish their metadata to metadata repositories that can then be queried to discover resources and services that have specified characteristics. Meta-data exists at all levels of the Grid from the lowest level repositories of Grid handles to the upper levels defining ontologies and other information about application resources.

Grid architectures are under very active and broad research and development around the World. They are in an early stage of evolution. It is hard to predict what approach, what architecture, what specific technology will finally win. Now it became apparent that the service-oriented paradigm provided the flexibility required for the third generation Grid. The Open Grid Services Architecture (OGSA) Framework, the Globus-IBM vision for the convergence of Web services and Grid computing, is a step in this direction. OGSA supports the creation, maintenance, and application of ensembles of services maintained by Virtual Organizations. It tailors the Web Services

approach to meet some grid-specific requirements. For example, OGSA supports interfaces for Service Discovery, Dynamic service creation, Lifetime management, Notification, Manageability, etc. Global Grid Forum (GGF) is the organizational unit that contributes to the development of OGSA.

Russian scientists take part in several international Grid-oriented projects. For example, in the European project “Enabling Grids for E-science in Europe” (EGEE) the following Russian organizations participate: Institute of High Energy Physics, Institute of Mathematical Problems of Biology of Russian Academy of Sciences, Institute of Theoretical and Experimental Physics, Joint Institute for Nuclear Research, Keldysh Institute of Applied Mathematics of Russian Academy of Sciences Moscow, Petersburg Nuclear Physics Institute of Russian Academy of Sciences, Russian Research Centre "Kurchatov Institute", Skobeltsyn Institute of Nuclear Physics of Moscow State University. They form one of the 12 federations of the EGEE project. Russian astronomers do not participate in the EGEE.

6.4 OGSA-DAI Architecture

Data Access and Integration is addressed within the Grid community primarily by the DAIS WG of the GGF [GGFOR], and by the OGSA-DAI project [OGDAI], which is producing the first reference implementation of data access and integration services; essentially those specified by the DAIS WG, although there are currently some terminological mismatches between the two. Taken together this activity represents the first concerted effort towards producing a generic framework for the integration of data access and computation. It aims to use the Grid to take specific classes of computation closer to the data, and to do this through the production of kits of parts for building tailored data access and integration applications. The initial focus is on (relational and XML) database access, but the conceptual vision extends to encompass all ways in which people store data.

This conceptual model is based on an external universe in which there are external data resource managers (e.g. DBMSs), external data resources (e.g. individual databases) and external data sets (e.g. query result sets extracted from a database). DAI Service classes (data resource manager, data resource, data activity session, data request and data set) then map onto these entities and mediate the interactions between them. The initial suite of DAI services are envisaged to run as follows [VODGR]:

1. A Client sends a request to a Registry, asking for sources of data relevant to “X”;
2. The Registry responds with a handle to a Factory;
3. The Client uses the handle to send the Factory a request for access to a database;

4. The Factory creates a GridDataService to manage that database access;
5. The Factory returns a handle to that GridDataService to Client;
6. The Client sends a query (in SQL, Xpath, etc) to the GridDataService;
7. The GridDataService interacts with the database to have the query executed;
8. The GridDataService returns the query results to the Client in XML.

All these stages are mediated by the exchange of messages over SOAP/HTTP, except for the interaction between the GridDataService and the database (which proceeds according to the database API) and the creation of the GridDataService by the Factory that produces an instantiation and obtains a handle to the instantiated object. Future DAI services are intended to embed application code into the GridDataService, and to allow the composition of more complicated services via the chaining of GridDataServices and GridDataTransformationServices.

The current OGSA-DAI product is represented by Release 5.0, which is built on top of the Globus [GLOBU] Toolkit 3.2.1 (http://www.ogsadai.org.uk/). R5.0 is primarily concerned with making changes to the various interfaces and document schema that will allow to support WS and WSRF based interfaces to the underlying data access and integration code. Still the finalisation of the Data Access and Integration Services (DAIS) standardization proposal is expected. Moving towards releasing the Distributed Query Processing software in a more closely integrated release is anticipated. The release 6.0 is scheduled for April 2005 [OGDA5]. It is expected that this release will provide a WSRF compliant interface to the OGSA-DAI functionality. It is also anticipated that the OGSA-DAI product will be available as an integrated part of the GT4.0 final release and as an independent package of services and APIs (which may bundle GT4.0 as part of the installation). The major features of this release include OGSA-DQP as an integrated part of release, fully compliant JDBC Driver for OGSA-DAI, support for WS-Security implementations, support for stored procedures on all supported databases (where available), SQL translation between vendor dialects for subset of queries, support for XQuery data resources.

OGSA-DAI components are either data access components or data integration components. A *Distributed Query Processing* (DQP) system is an example of a data integration component and can potentially provide effective declarative support for service orchestration as well as data integration. The service-based DQP framework provides an approach that:

- supports queries over GDSs and over other services available on the Grid, thereby combining data access with analysis;

- uses the facilities of the OGSA to dynamically obtain the resources necessary for efficient evaluation of a distributed query;
- adapts techniques from parallel databases to provide implicit parallelism for complex data-intensive requests;
- uses the emerging standard for GDSs to provide consistent access to database metadata and to interact with databases on the Grid.

OGSA-DQP provides two services to fulfill its functions: *The Grid Distributed Query Service (GDQS)* and *the Grid Query Evaluation Service (GQES)*. The GDQS provides the primary interaction interfaces for the user, collects the necessary metadata and acts as a coordinator between the underlying query compiler/optimizer engine and the GQES instances. GQES instances are created and scheduled dynamically, to evaluate the partitions of a query constructed by the optimizer of the GDQS.

6.5 WSRF

WSRF (Web Services Resource Framework, http://www.globus.org/wsrf) is Web services for grid computing. WSRF defines conventions for managing 'state' so that applications can reliably share changing information. In combination with WS-Notification and other WS- standards, the result is to make grid resources accessible within a web services architecture. Coupled with WS-Notification, the specification is a response to, and supersedes, the grid community's own first effort to converge grid and web services, the Open Grid Service Infrastructure (OGSI), which the Global Grid Forum (GGF) and others released in 2003. Announced by the Globus Alliance and IBM (with contributions from HP, SAP, Akamai, Tibco and Sonic) in January 2004, WSRF is due to be implemented in version 4.0 of the open source Globus Toolkit for grid computing, as well as several commercial packages. It consists of several component specifications, including WS-Resource Properties, WS-ResourceLifetime, WS-ServiceGroup and WS-BaseFaults.

7 Classes of astrophysical problems for VO

7.1 Class of problems solvable applying database search technique

A Virtual Observatory (VO) is a collection of interoperating data resources and software tools which utilize the internet to form a scientific research environment in which astronomical research programs can be conducted. In VO large surveys and catalogues should be joined into a uniform and interoperating "digital universe" providing for entirely new areas of astronomical research. Astronomical data falls into two broad categories: catalog (hundreds of attributes for billions of objects) and image (10s TB of pixel data). The specific

data classes include source catalog, time series, event list, visibility data, (including the various image subclasses), spectrum.

For VO as an integrated collection of observational data, the class of astronomical problems that can be investigated applying database search technique is the basic one. Remarkable sample of such problems has been provided for the Sloan Digital Sky Survey (SDSS) – a 5-wavelength catalog over 10,000 square degrees of the sky. Data acquisition and archiving for SDSS has been designed for online interactive analysis. The 200 million objects in the multi-terabyte database have mostly numerical attributes in a 100+ dimensional space. Points in this space have highly correlated distributions. The archive enables astronomers to explore the data interactively. Data access is aided by multidimensional spatial and attribute indices.

The following query examples [SSDSS] reveal respective classes of problems that can be explored interactively. The queries correspond to typical tasks astronomers would do with a C++ program, extracting data from the archive, and then analyzing it. Being able to state queries simply and quickly could be a real productivity gain for the Astronomy community. These tasks have fairly simple SQL equivalents. Often the query can be expressed as a single SQL statement. In some cases, the query is iterative, the results of one query feeds into the next. Representation of the queries in SQL and results of their implementation can be found in [GSDSS].

Q1: Find all galaxies with unsaturated pixels within 1 arcsecond of a given point in the sky (right ascension and declination). This is a classic spatial lookup. A quad-tree spherical triangle index with object type (star, galaxy, etc.) as the first key and then the spatial attributes are assumed.

Q2: Find all galaxies with blue surface brightness between 23 and 25 mag per square arcseconds, and – 10 < super galactic latitude (sgb) < 10, and declination less than zero. This searches for all galaxies in a certain region of the sky with a specified brightness in the blue spectral band. The query uses a different coordinate system, which must first be converted to the hierarchical triangles. It is then a set of disjoint table scans, each having a compound simple predicate representing the spatial boundary conditions and surface brightness test.

Q3: Find all galaxies brighter than magnitude 22, where the local extinction is > 0.75. The local extinction is a map of the sky telling how much dust is in that direction, and hence how much light is absorbed by that dust. The extinction grid is stored as a table with one square arcminute resolution – about half a billion cells. The query is either a spatial join of bright galaxies with the extinction grid table, or the extinction is stored as an attribute of each object so that this is just a scan of the galaxies in the Photo table.

Q4: Find galaxies with a surface brightness greater than 24 with a major axis 30"<d<1', in the red-band, and with an ellipticity > 0.5. Each of the 5 color

bands of a galaxy have been pre-processed into a bitmap image which is broken into 15 concentric rings. The rings are further divided into octants. The intensity of the light in each ring is analyzed and recorded as a 5x15 array. The array is stored as an object (SQL blob in our type impoverished case). The concentric rings are pre-processed to compute surface brightness, ellipticity, major axis, and other attributes. Consequently, this query is a scan of the galaxies with predicates on precomputed properties.

Q5: Find all galaxies with a deVaucouleours profile (r¼ falloff of intensity on disk) and the photometric colors consistent with an elliptical galaxy. The deVaucouleours profile information is precomputed from the concentric rings as discussed in Q4. This query is a scan of galaxies in the Photo table with predicates on the intensity profile and color limits.

Q6: Find galaxies that are blended with a star, output the deblended magnitudes. Preprocessing separates objects that overlap or are related (a binary star for example). This process is called deblending and produces a tree of objects; each with its own 'deblended' attributes such as color and intensity. The parent child relationships are represented in SQL as foreign keys. The query is a join of the deblended galaxies in the photo table, with their siblings. If one of the siblings is a star, the galaxy's identity and magnitude is added to the answer set.

Q7: Provide a list of star-like objects that are 1% rare for the 5-color attributes. This involves classification of the attribute set and then a scan to find objects with attributes close to that of a star that occur in rare categories.

Q8: Find all objects with spectra unclassified. This is a sequential scan returning all objects with a certain precomputed flag set.

Q9: Find quasars with a line width > 2000 km/s and 2.5<redshift<2.7. This is a sequential scan of quasars in the Spectro table with a predicate on the redshift and line width. The Spectro table has about 1.5 million objects having a known spectrum but there are only 100,00 known quasars.

Q10: Find galaxies with spectra that have an equivalent width in Ha > 40 Å (Ha is the main hydrogen spectral line). This is a join of the galaxies in the Spectra table and their lines in the Lines table.

Q11: Find all elliptical galaxies with spectra that have an anomalous emission line. This is a sequential scan of galaxies (they are indexed) that have ellipticity above .7 (a precomputed value) with emission lines that have been flagged as strange (again a precomputed value).

Q12: Create a grided count of galaxies with u-g>1 and r<21.5 over 60<declination<70, and 200<right ascension<210, on a grid of 2', and create a map of masks over the same grid. Scan the table for galaxies and group them in cells 2 arc-minutes on a side. Provide predicates for the color restrictions on u-g and r and to limit the search to the portion of the sky defined by the right ascension and declination conditions. Return the count of qualifying galaxies in

each cell. Run another query with the same grouping, but with a predicate to include only objects such as satellites, planets, and airplanes that obscure the cell. The second query returns a list of cell coordinates that serve as a mask for the first query. The mask may be stored in a temporary table and joined with the first query.

Q13: Create a count of galaxies for each of the HTM triangles (hierarchal triangular mesh) which satisfy a certain color cut, like 0.7u-0.5g-0.2 and i-magr<1.25 and rmag<21.75, output it in a form adequate for visualization. This query is a sequential scan of galaxies with predicates for the color magnitude. It groups the results by a specified level in the HTM hierarchy (obtained by shifting the HTM key) and returns a count of galaxies in each triangle together with the key of the triangle.

Q14: Provide a list of stars with multiple epoch measurements, which have light variations >0.1 magnitude. Scan for stars that have a secondary object (observed at a different time) with a predicate for the light variations.

Q15: Provide a list of moving objects consistent with an asteroid. Objects are classified as moving and indeed have 5 successive observations from the 5 color bands. So this is a select of the form: select moving object where sqrt((deltax5-deltax1)2 + (deltay5-deltay1)2) < 2 arc seconds.

Q16: Find all star-like objects within DeltaMagnitde of 0.2 of the colors of a quasar at 5.5<redshift<6.5. Scan all objects with a predicate to identify star-like objects and another predicate to specify a region in color space within 'distance' 0.2 of the colors of the indicated quasar (the quasar colors are known).

Q17: Find binary stars where at least one of them has the colors of a white dwarf. Scan the Photo table for stars with white dwarf colors that are a child of a binary star. Return a list of unique binary star identifiers.

Q18: Find all objects within 1' of one another other that have very similar colors: that is where the color ratios u-g, g-r, r-I are less than 0.05m (Magnitudes are logarithms so these are ratios). This is a gravitational lens query. Scan for objects in the Photo table and compare them to all objects within one arcminute of the object. If the color ratios match, this is a candidate object. We may precompute the five nearest neighbors of each object to speed up queries like this.

Q19: Find quasars with a broad absorption line in their spectra and at least one galaxy within 10". Return both the quasars and the galaxies. Scan for quasars with a predicate for a broad absorption line and use them in a spatial join with galaxies that are within 10 arc-seconds. The nearest neighbors may be precomputed which makes this a regular join.

Q20: For a galaxy in the BCG data set (brightest color galaxy), in 160<right ascension<170, 25<declination<35, give a count of galaxies within 30" which have a photoz within 0.05 of the BCG. First form the BCG (brightest galaxy in

a cluster) table. Then scan for galaxies in clusters (the cluster is their parent object) with a predicate to limit the region of the sky. For each galaxy, test with a sub-query that no other galaxy in the same cluster is brighter. Then do a spatial join of this table with the galaxies to return the desired counts.

7.2 Classes of general problems for VO

A list of problems have been registered for the AstroGrid project (http://wiki.astrogrid.org/bin/view/VO/ScienceProblemList). The problems in the list are provided with a brief description that looks as it is shown for the Galaxy Clustering:

Science Problem*: GalaxyClustering*

Science Goal: Aim to address the evolution of galaxy populations in clusters. Comparison can then be made with n-body model data.

ProblemDescription: Clusters of galaxies can be used to trace distribution of matter in the universe over large scales. Clusters are typically X-ray (see e.g. http://wiki.astrogrid.org/bin/view/VO/XrayGalaxyClusterSurvey) or optically selected. Many optically selected cluster samples have suffered from various selection effects – such as the use of only one color data (e.g. http://wiki.astrogrid.org/bin/view/VO/GalaxyClustering #JumpToDalton1992).

New techniques (e.g. http://wiki.astrogrid.org/bin/view/VO/ GalaxyClustering#JumpToGal2000) select clusters using multicolor data to localize clusters which are predicted to contain an overabundance of red, early type galaxies. Cluster identification uses positional information to select clusters (e.g. http://wiki.astrogrid.org/bin/view/VO/GalaxyClustering #JumpToGladders2000)

Cluster distributions can be compared to matter distributions generated by e.g. Lambda CDM models (e.g. http://wiki.astrogrid.org/bin/view/VO/ GalaxyClustering#JumpToNagamine2001) or Warm Dark Matter models (e.g. http://wiki.astrogrid.org/bin/view/VO/GalaxyClustering#JumpToBode2001). These models now have sufficient resolution to show dwarf galaxies.

CurrentSolution: Procedure is to select sources marked as galaxies, select only those in a particular locus of the (g-r) vs (i-r) color space, and then create density maps (see e.g. Gal et al, 2000).

Redshifts are determined photomterically and spectrscopically. For clusters found, a serach on cluster members would be required cross referenced against possible spectroscopic data on galaxies in those clusters to fix the spectrscopic redshift to the galaxy. This could require searching information contained within e.g. NED (http://nedwww.ipac.caltech.edu/).

VOSolution: Application to deeper surveys would allow clusters at higher redshifts to be located. Application to surveys with more colors (e.g. Opt+IR) would enable higher precision cluster selection.

Large n-body code model outputs will need to be compared with real observed cluster distributions. Issues include interfacing to large model data sets, visualization of model vs real data – e.g. matter vs clusters at ranges of redshift, statistical correlations etc.

Roughly the problems in the AstroGrid list can be classified as follows (each problem in the list has respective definition at (http://wiki.astrogrid.org/bin/view/VO/ScienceProblemList):

1. Cosmology;
    - NonCosmologicalRedshifts;
    - EvolutionOfBias;
    - Cosmic microwave background (NVO);
    - Virtual Universe (UK);
2. Galaxy formation and development;
    - ActiveGalaxiesUnbiasedStudy;
    - GalacticMaserFaradayRotation;
    - GalaxyMorphologyRecognition;
    - Galaxy correlation functions (NVO);
    - GalaxyClustering;
    - Dynamical state of galaxy clusters (NVO);
    - GalaxyMorphologyProbingStarFormation;
    - GalaxySpectralAnalysis;
    - GetGalaxyRedshift;
    - LowSurfaceBrightnessGalaxyDiscovery;
    - XrayGalaxyClusterSurvey;
    - OpticalNearIRGalaxyClusterSelection;
    - QuasarHostGalaxiesBlackHoleMasses;
    - HiZQuasars;
    - MultiVariateGalaxyProperties;
3. Star formation and evolution;
    - AGBstarCandidates;
    - ActiveStarProperMotions;
    - DynamicsofAGNandStarburstNucle;
    - StarFormingRegionStructure;
    - UnusualStarsInGalacticPlane;
    - BrownDwarfSelection;

- SupernovaGalaxyEnvironment;
- HaloWhiteDwarfs;

4. Sun and Planets;
   - SolarAbundances;
   - SolarCoronalWaveHeating;
   - SolarCoronalWaves;
   - SolarFlareOnset;
   - SolarStellarFlareComparison;
   - SolarSystemObjects;
   - SolarVicinityStars;
   - STPSolarEventCoincidence;
   - CoronalMassEjectionEffect;
   - CoronalMassEjectionTrigger;
   - BoundaryLayerStructureAndStability;
   - MagneticStormOnset.

AstroGrid attempts also to define use cases that might be useful for the problem solving. Use cases are defined according to the following pattern:

UseCase: SyntheticSpectra

EndResult: Spectrum of desired object(s) is (are) returned covering the broadest wavelength range for which data is available

PreConditions: Input list of objects with positional (plus position error) information.

FlowOfEvents:

User has list of objects for which the spectral energy distribution is desired over a complete a wavelength range as possible.

User is returned a list of all data that could match query, taking into account position, error, size of target source, size of aperture on sky of the instrument.

User selects data sets of interest, or accepts default best set of data, or accepts all possible data.

Data is extracted and coverted to a common flux scale.

Data for which no flux information is available is normalized to 'best fluxed' data.

Data is returned in the form of an energy vs wavelength distribution, with an indication of the provenance of each data point.

PostCondition:

All data will be on a common flux scale. It should be possible to zoom in on certain wavelength regions where data of high spectral resolution might be available.

BasicAssumptions:

Both spectral and photometric data points should be gathered.

Data flagged as of photometric quality should be used when normalizing the fluxes of non-photomtric or un-fluxed data.

"Extracts spectral plots" in conjunction with "browses" suggests that there's some general browsing tool that (a) can detect and correlate records that match by position and (b) knows how to do the plots. This won't work well unless each survey returns its spectophotometry in a standard format. If the optical results are in magnitudes, the radio results in janskies and the X-ray results in counts per square arcsecond per second for survey X and Wm-2 for survey Y, then correlating the data will be hard.

A list of AstroGrid Use cases (http://wiki.astrogrid.org/bin /view/VO/UseCases) arranged into a simple classification follows:

1. Managerial;
    - UseCase: AuthenticateIdentity;
    - UseCase: NegotiateAccessToResource;
    - UseCase: DetermineAuthority;
    - UseCase: AccessLibraryFunction;
    - UseCase: AccessWebService;
    - UseCase: PerformRegistrySearch;
2. General;
    - UseCase: AstrometryBootstrap: List of astrometric positions, either as accurate as possible or to within specified tolerance;
    - UseCase: BulkSpectralLineMeasures: Measurement of Equivalent Widths of selected lines of bulk selected candidates;
    - UseCase: GetFluxOrUpperLimitAtPosition: Astronomer gets flux or upper limit in given band at particular source position;
    - UseCase: GetReducedSpectra: Astronomer wants to get actual spectra (as opposed to data derived from spectra like spectral indices) for a given spectral range and taken near a given position;
    - UseCase: RedshiftSpectralDetermination;
    - UseCase: DataQuality: User receives data quality information or "no data available" message for selected dataset;
3. Special;

- UseCase: StarRecognition: List of objects which are stars according to defined criteria to within a certain probability;
- UseCase: ActiveStarID: A list of active stars for correlation with other parameters;
- UseCase: ActiveStarsforProperMotion: A list of radio stars which are close enough so that, if proper motions are measured using VLBI, this could reveal reflex motion due to sub-stellar companions;
- UseCase: BrownDwarfRecognition;
- UseCase: GalaxyMorphologyRecognition: Classification of galaxies according to the probability that they are (non-exclusively): Spiral (and subclasses), Elliptical (and subclasses), AGN (etc), Starburst Irregular.

7.3 Theoretical research and VO

According to the British Virtual Universe proposal (at http://star-www.dur.ac.uk/~csf/virtU/virtU-final.pdf) it is planned to construct a Virtual Universe (VirtU) consisting of the "Theoretical Virtual Observatory" (TVO) and associated diagnostic tools making up the "Theory/Observations Interface" (TOI). VirtU [VIRTU] will be built upon a dynamic archive constructed, in the first instance, from state-of-the-art cosmological simulations which encode our current understanding of our world model, of the clustering evolution of dark matter and of the physics of galaxy formation. By including the latest, most realistic models of galaxies, clusters and other structures, together with relevant analysis tools, this virtual universe will be a powerful entity providing a service to a wide community of theorists, phenomenologists, observers and instrumentalists. VirtU, however, is not restricted to Cosmology. In the longer term, it will encompass a wider range of simulations, covering, for example, star and planet formation. VirtU is the response of a large community of theorists, phenomenologists, observers and instrumentalists to the challenges presented by the ongoing explosion of observational and simulated data. VirtU is an essential complement to AstroGrid and will form a key component of the Virtual Observatory (VO). VirtU has two major components, the "Theoretical Virtual Observatory" (TVO), and a toolkit of applications targeted at the theory/observations interface (TOI). Fig. 3 shows the relationships between the TVO, TOI and AstroGrid.

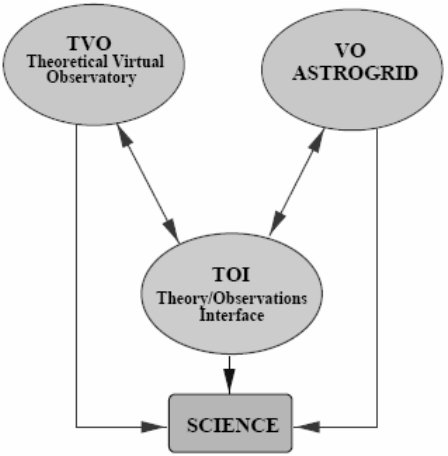

Figure 3. The relationship between the TVO, TOI and AstroGrid

To exploit the scientific opportunities offered by advances in simulation capability and in the quality of observational data, it is necessary to develop an interface to facilitate the comparison of models with data. This must be responsive to the limitations of both models and data and attempt to bring both on to a common plane by, for example, including noise or selection effects in the models or degrading the resolution of an observed image to the resolution of a simulation. TOI will develop this interface and make available on the grid a suite of diagnostics that include traditional as well as novel approaches to data compression, classification and parameter estimation in large datasets of galaxy spectra and images. Fig. 4 illustrates the VirtU elements interconnection in a schematic way.

The VirtU model, now widely accepted as the standard cosmogony, is based on two key assumptions: (i) that the Universe underwent an early period of inflationary expansion during which its curvature was flattened and small irregularities of quantum origin were imprinted and (ii) that these irregularities grew into cosmological structures by gravitational evolution driven by massive, weakly interacting elementary particles or cold dark matter (CDM). This model agrees with the distribution of galaxies, as mapped by the new generation of surveys. In the first instance, the TVO will be built around the Millennium simulation, the largest and most ambitious cosmological simulation ever conceived. The Millennium simulation will follow 10^{10} particles in a universe 500 Mpc on a side.

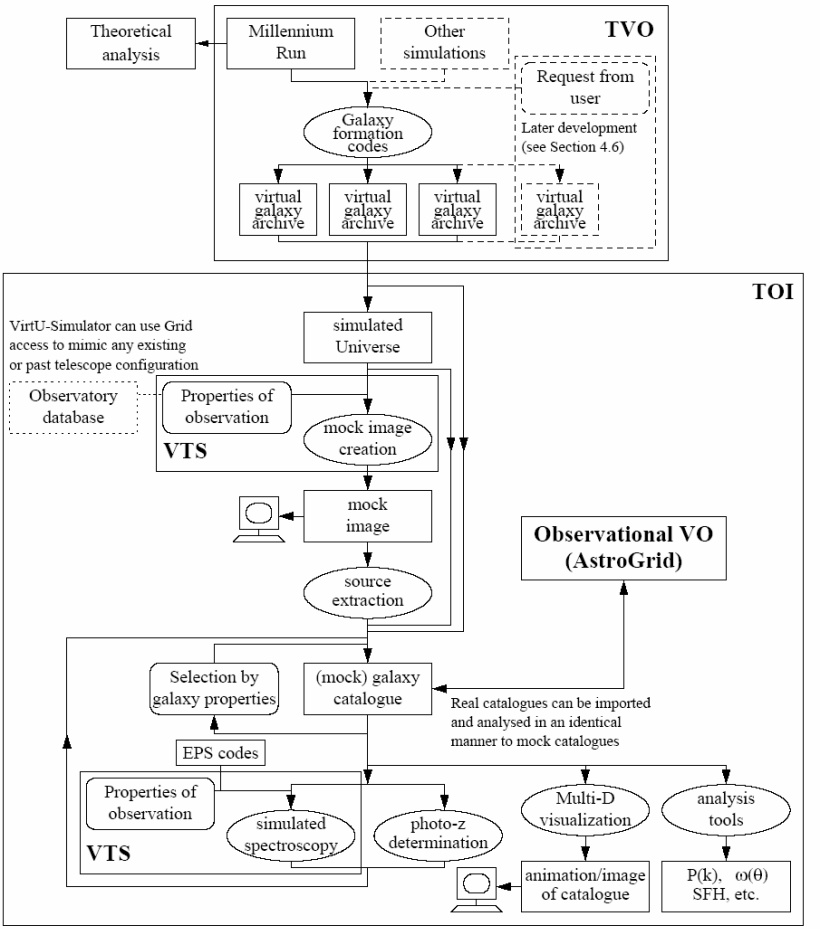

Figure 4. Example of extragalactic application of VirtU

This simulation will track the clustering evolution of dark matter and the formation history of about 50 million galaxies. Data will be output about 50 times (enough to construct lightcones and make movies) and each output will consist of 300 GByte. Thus, the raw data alone will amount to 15 Tbytes. The simulation is planned to be performed on the IBM-Power4 768-processor Regatta of the Max-Planck Rechenzentrum in Garching. This supercomputer can provide the required 1 Tbyte of memory.

7.4 Co-existence of theoretical and observational archives and services in VO

One of the main goals of the TVO is to enable *federation* of theoretical with observational archives [LTHEO]. That is, to enable the comparison of observational with equivalent theoretical data products in a uniform manner. An example is the comparison of synthetic galaxy catalogues, such as those arising from semi-analytical algorithms applied to dark matter simulations, with observational catalogues from the SDSS. Another example is color-magnitude diagrams for globular clusters, both observed and simulated. Thus, emphasis will be on put the publication of these products and the construction of services (algorithms, visualization tools, etc.) that enable this kind of comparison. Obviously, this ties in directly with the needs of observers. At the same time, it automates what theoreticians have to do when they want to estimate the validity of their models. Standard observational data products that have so far been dealt with explicitly in the VO are images, spectra, and source catalogues. It is planned that also time-ordered event lists and radio visibility data will be supported. Likewise it is required to identify some standard theoretical data products. This may help decide on which standard services to define for the TVO. Examples of products that are of particular use to the theory/observational interface, include:

- synthetic observations of X-Ray clusters vs. XMM/Chandra observations;
- color-magnitude diagrams of globular clusters observed vs. simulated;
- galaxy catalogues from semi-analytical work vs. observations (for example SDSS);
- galaxy merger simulations vs. observations;
- Planck CMB simulations with non-trivial topologies.

A representative set of *classes* of theoretical models/simulations that people are interested in is to be identified. It is the task of the IVOA data-modeling group to describe these in a way that extracts the common elements, but also allows for the differences. Several ways of initial simulations classifications (namely by simulation subject, by simulated physical processes, by the software algorithms used in these simulations, and by the produced types of data products) can be considered. For instance, classification by *subject* of simulations looks as follows:

- CMB;
- large-scale structure analysis: gravitational lensing, Lyman alpha cloud spectra, pencil beams, semi-analytical galaxy formation, gravitational clustering, clusters;
- galaxy clusters;

- galaxy formation;
- galaxy mergers;
- globular cluster;
- molecular clouds;
- stellar evolution tracks;
- supernovae;
- accretion disks;
- gravitational waves from merging black holes;
- planetary systems;
- spectra;
- jets.

Three general classes of services are considered:

Query and browsing services are aimed at the discovery of specific data products in an archive based upon specific query parameters. Experience from the VO efforts so far indicates that it will probably be useful to define some simple, standard query services that are easily implemented by simulation archives. Examples from the observational VO are the Simple Cone Search, the Simple Image Access Protocol, and the Simple Spectrum Access Protocol. From TVO point, it is required to define some alternative query services that are relevant. An example might be to return all particles from a cosmological N-body simulation within a given sized volume randomly positioned in space at a given redshift.

An analysis service is defined as a software component that performs a manipulation of data to extract new information. This is also often called *data mining* meaning here discovery of correlations that are *not* explicitly modeled in a schema. The most important standard example from the observational VO is a cross-match service, which identifies common objects in different source catalogues. Examples are:

- Virtual (or synthetic) telescope. Service that “observes” simulation results to produce “images” that can be directly compared to observations;
- Comparators for comparing the results of these synthetic telescopes to the actual observations;
- Statistics calculators such as n-point functions, morphology indicators etc;
- Halo finders;
- Visualization services.

Simulations through the VO. We here extend the definition of "simulation" to include every algorithm that creates *new* data, possibly from data products that have been published in the VO already. The distinction between this and the analysis/data mining services of the previous section is somewhat blurry, though. Some proposals are:

- N-body codes for galaxy mergers;
- Semi-analytical galaxy formation algorithms on halo-merger trees;
- N-body codes linked to stellar evolution codes for globular cluster simulations.

Specificity of TVO (http://www.ivoa.net/twiki/bin/view/IVOA/IvoaTheory).

The following specific features of TVO vs observations are emphasized [LTHEO]:

1. Position based query protocols irrelevant for TVO. Simple query protocols (such as the Simple Cone Search and the Simple Image Access Protocol) are based on absolute positions on the sky. Results of simulations usually are not tied to a specific absolute position on the sky. TVO protocols will need to query for different properties rather than spatial location;
2. Matching based on similarity, not identity. Crossmatch based on identity is irrelevant for theory data products. Finding a "match" in a theoretical archive for a source observed in some catalogue will not be based on closeness in position, but on closeness in physical parameter space. Parameters on which theoretical archives will be searched will consist of *physical properties*, such as masses or sizes of simulated objects, or temperatures or luminosities;
3. New observables. In observations results are almost invariably based on detected photons. In theory this generally is not the case. Entities and properties that might not be directly observable are commonly used for theoretical models and are thus included in the results. New theoretical concepts/entities are to be identified;
4. Simulated vs observational properties. Theoretical data products contain, by construction, knowledge of *all* quantities of interest. To compare such results to observational data, one may need to modify these quantities in a way that mimics observations;
5. Variety of models. Simulation products can be queried by the kind of physical model that underlies them, something that seems less relevant for observations. Different algorithms used for the same problem might lead to different ranges of validity of the result (e.g., different limitations as far as resolution, smallest resolved scales, largest resolved scales). These restrictions also have to be included in the data

model to allow for comparisons between different theoretical results and between theoretical results and observations;

6. Creating artificial observations from simulations. Comparisons between observations and theory will often involve the extraction of observation-like data products from theoretical ones, whereas the opposite is less likely to be required. One may want to store sets of *very commonly used derived data* together with the "raw" original data products.

Concrete tasks:

1. Archive publication and querying;
   - Define a conceptual data model for simulations compatible with the IVOA data model;
   - Create a reference implementation for a meta-data repository;
   - Design and implement automated registration services for meta-data repository;
   - Implement query services on repository;
   - Define standard queries/protocols for theory analogous to SCS/SIAP for observations ("volume of space at given cosmological time", possibly subsampled by certain factor; properties (constituents) of identified objects as function of time: synthetic spectra for specific galaxies/stars/...);
2. Analysis services;
   - Create webservice interfaces for existing services (such as halo finders, statistics calculators);
   - Create tools for creating mock observational products from simulations (Virtual Chandra for X-Ray clusters, Mock SDSS from semi-analytical galaxy catalogues, Weak lensing on LSS, Strong lensing on (X-Ray) clusters, Globular cluster observations, Mock spectra, Mock CMB maps);
   - Create tools for comparing observations with mock-observations theoretical products, for example (X-Ray clusters, Galaxy catalogues, Galaxy mergers, Globular cluster CMDs);
   - Expose visualization services;
3. Simulators;
   - initial conditions generator;
   - galaxy merger simulator;
   - all these have to be wrapped by a web interface;

- semi-analytical galaxy formation wrapped by web interface;

4. Reference implementations;
   - Semi-analytical galaxy formation;
   - X-Ray clusters simulation vs observation;
   - Galaxy merger simulation reproducing observation;
   - Advantages of storing theory data in a standard database: example, store simulated galaxy catalogues in SDSS, in same format.

8 Virtual observatory architecture according to IVOA

RVO infrastructure should be based on the IVOA standards. This section provides an overview of the IVOA standards in accordance with their state at the end of 2004. The development of architectural decisions and standards is accomplished by 8 IVOA Working Groups:

- Resource Registry: The IVOA Registry will allow an astronomer to be able to locate, get details of, and make use of, any resource located anywhere in the IVO space, i.e. in any Virtual Observatory. The IVOA will define the protocols and standards whereby different registry services are able to interoperate and thereby realize this goal.
- Data Modeling: defines the standardized data models describing the structure and semantics of astronomical data to permit dataset interoperability.
- Content Description (UCD): metadata definition and standardization based on the unified content descriptors (UCD).
- Data Access Layer: definition and formulation of VO standards for remote data access.
- VOTable: development of an XML standard for storage and interchange of data represented as a set of tables.
- VO Query Language: development of query language for distributed astronomical resources.
- Grid & Web Services: use of Grid technologies and Web Services in the VO context, and investigation of required standards in this area.
- Standards & Processes: development the standards and documents acceptance process in IVOA (http://www.ivoa.net/twiki/bin/view/IVOA/IvoaStdsDocsProc).

The IVOA defined a process by which official IVOA standards documents are advanced from Working Drafts to formal Recommendations:

- IVOA Notes: Dated, public record of an idea, comment, or document;
- IVOA Working Drafts: published at the discretion of Working Groups, they are subject to review by the Document Coordinator for compliance to the guidelines;
- IVOA Proposed Recommendation: published by the chair of a Working Group, they are considered to be technically mature and ready for wide review;
- IVOA Recommendation: published by the IVOA Executive Committee, they are the final form of IVOA documents and constitute an IVOA Standard.

8.1 VO architecture overview

The architecture of the VO (http://www.ivoa.net/twiki/bin/view /IVOA/IvoaArchitecture) is *service oriented*, meaning that components of the system are defined by the nature of requests and responses to services. Because of this, the description of each service is based on the choice of the protocols for requests and responses, rather than classes and methods. Data is communicated between services in two basic formats: FITS and XML.

Fig. 5 provides a high-level conceptual overview of the IVOA-supported architecture of Virtual Observatories that has emerged over the last few years.The top bar of the figure represents this objective: discovery of data and services, reframing and analyzing that data through computation, publishing and dissemination of results, and increasing scientific output through collaboration and federation. The IVOA does not specify or recommend any specific portal or library by which users can access VO data, but some examples of these portals and tools are shown in the grey box.

Different colored vertical arrows represent the different service types and XML formats by which these portals interface to the IVOA-compliant services. In the IVOA architecture, we have divided the available services into three broad classes:

- Data Services, for relatively simple services that provide access to data;
- Compute Services, where the emphasis is on computation and federation of data;
- Registry Services, to allow services and other entities to be published and discovered.

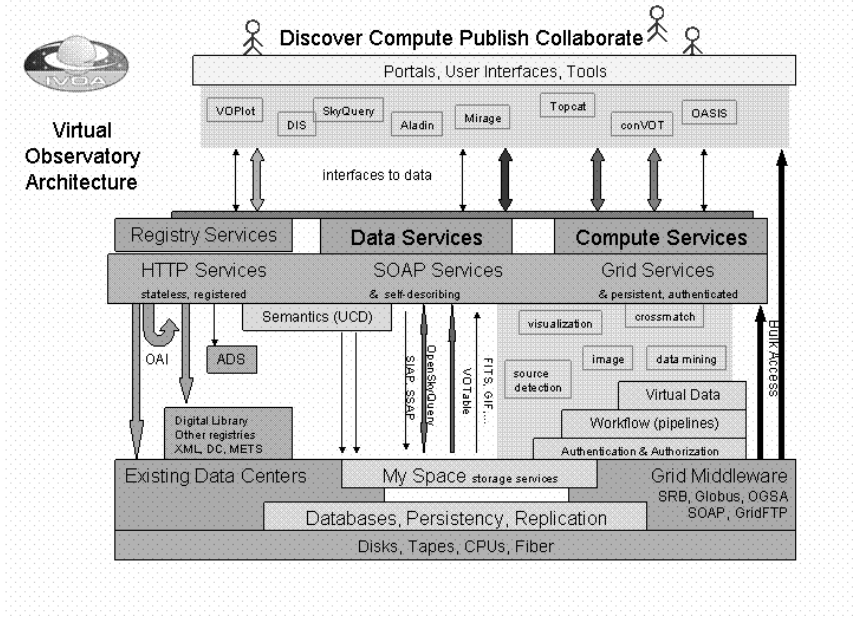

Figure 5. IVOA architecture

These services are implemented at various levels of sophistication, from a stateless, text-based request-response, up to an authenticated, self-describing service that uses high-performance computing to build a structured response from a structured request. In the VO, it is intended that services can be used not just individually, but also concatenated in a distributed *workflow*, where the output of one is the input of another. The registry services facilitate publication and discovery of services.

Each registry has three kinds of interface: publish, query, and harvest. People can publish to a registry by filling in web forms in a web portal, thereby defining services, data collections, projects, organizations, and other entities. The registry may also accept queries in a one or more languages, and thereby discover entities that satisfy the specified criteria. The third interface, harvesting, allows registries to exchange information between themselves, so that a query executes at one registry may discover a resource that was published at another. Registry services expect to label each VO resource through a universal identifier, that can be recognized by the initial string ivo://. Resources can contain links to related resources, as well as external links to the literature, especially to the Astronomical Data System. The IVOA registry architecture is compliant with digital library standards for metadata harvesting and metadata schema, with the intention that IVOA-compliant resources can appear as part of every University library.

Data services range from simple to sophisticated, and return tabular, image, or other data. At the simplest level (conesearch), the request is a cone on the sky (direction/angular radius), and the response is a list of "objects" each of which has a position that is within the cone. Similar services (SIAP, SSAP) can return images and spectra associated with sky regions, and these services may also be able to query on other parameters of the objects.

The OpenSkyQuery protocol drives a data service that allows querying of a relational database or a federation of databases. In this case, the request is written in a specific XML abstraction of SQL that is part of ADQL (Astronomical Data Query Language).

The IVOA architecture will also support queries written at a more semantic level, including queries to the registry and through data services. To achieve this, the IVOA is developing a structured vocabulary called UCD (Unified Content Descriptor) to define the *semantic type* of a quantity.

The IVOA expects to develop standards for more sophisticated services, for example for federating and mining catalogs, image processing and source detection, spectral analysis, and visualization of complex datasets. These services will be implemented in terms of industry-standard mechanisms, working in collaboration with the grid community.

Members of the IVOA are collaborating with a number of IT groups that are developing workflow software, meaning a linked set of distributed services with a dataflow paradigm. The objective is to reuse component services to build complex applications, where the services are insulated from each other through well-defined protocols, and therefore easier to maintain and debug. IVOA members also expect to use such workflows in the context of *virtual data*, meaning a data product that is dynamically generated only when it is needed, and yet a cache of precomputed data can be used when relevant.

In the Fig. 5, the lowest layer is the actual hardware, but above that are the existing data centers, that implement and/or deploy IVOA standard services. Grid middleware is used for high-performance computing, data transfer, authentication, and service environments. Other software components include relational databases, services to replicate frequently used collections, and data grids to manage distributed collections.

A vital part of the IVOA architecture is *MySpace* so that users can store data within the VO. MySpace stores files and DB tables between operations on services; it avoids the need to recover results to the desktop for storage or to keep them inside the service that generated them. Using MySpace estblishes access rights and privacy over intermediate results and allows users to manage their storage remotely.

The IVOA architecture uses services at different levels: HTTP GET/POST services, SOAP services, Grid services. In the IVOA architecture, a VO-compliant web service is defined as one that can also supply a VOResource

description of the service, including curation, description, sky region, IVOA identifier, and other information.

8.2 Data Modeling

8.2.1 A unified domain model for astronomy, for use in the Virtual Observatory

The document “A unified domain model for astronomy” (http://www.ivoa.net/twiki/bin/view/IVOA/IvoaDataModel) is the IVOA attempt to define a conceptual model, created as the result of the domain knowledge extracted from the modelers (some of whom are astronomers) and their direct coworkers as well as from literature and other external references. Authors consider that the model can be used in various ways:

- It can be used as the basis for a meta-data repository that archives can use to describe their data products in a common model;
- It can be used as a model describing the entities (classes and attributes) that can be used in a common query language for these astronomical archives and for the relations between that can be followed from these entities in navigation to related ones;
- It can be mapped to an XML schema, to a Java or C# class library, to a relational database schema, allowing reference implementations for these particular bindings;
- It can simply serve as a formal, common language in “whiteboard discussions” about the structure of particular data products.

Technically the domain model is defined using the UML class diagrams. Operations in class definitions in the document "A unified domain model for astronomy" are avoided. Actually this makes the definition looking not as truly object-oriented.

Though for RVO the approach looks attractive (domain model might be used as a mediator schema), it is doubtful that global data model for the whole domain of astronomy could succeed. Each class of astronomical problems will introduce its own concepts, data structures, behaviors convenient for the respective problems (see section 7 on the classes of astrophysical problems). Each new instrument and changing in observational technology will lead to new kinds of data that could not be foreseen in advance. Therefore, it seems that data modeling approach should provide much more flexibility to survive.

It is said in the document that the way of using the common domain model is equivalent to an *ontology*. At the same time the main difference between conceptual model and ontology consists in the following. Conceptual model can be used as a global schema over existing heterogeneous data sources and services. It means that existing sources/services can be registered at the

conceptual model, mapped to it so that querying through the domain definition of the registered sources could be possible. Ontology is used as a reference definition of the domain concepts and relationships between them. Such definitions of concepts can be used for annotation of elements of various data models in the domain to provide them with the adequate semantics (cf. UCDs as a step towards simple ontology).

What is defined now in the document "A unified domain model for astronomy" might be more suitable to consider as an attempt to provide a draft definition of an ontology for the domain as a description of sets of concepts and relationships between them.

Alongside with a unified domain model, specific data models are being defined for various kinds of astronomical data, such as Spectra, Quantity, Observations, Transforms, Catalogs, Inteferometry, Simulations, Passband, Error/Accuracy. Some of these models are overviewed in the subsequent subsections.

8.2.2 Data model for quantity

A VO data model to describe the semantic content of sets of astronomical data values and their most closely associated metadata has been defined (http://www.ivoa.net/twiki/bin/view/IVOA/IvoaDataModel). The model may be used by aggregation or extension in higher level models describing astronomical datasets. Any value must be associated with a physical concept which can be tagged as a UCD, and with a physical unit. A set of interfaces to an object called Quantity and to some related objects is intended to be defined.

The following concepts are involved into the quantity model: accuracy, quality, array axes, coordinates, frames, coordinate systems, units, transformations. Interfaces denoted BasicQuantity, CoreQuantity, StandardQuantity, Unit, DataType, Region, Locator, Accuracy, Frame, CoordinateSystem are defined.

XML serialization for the proposed values is also defined.

8.2.3 IVOA Observation data model

A comprehensive data model named ’Observation’ for observational data is currently being defined (Data Model for Observation, IVOA WG internal draft). An Observation can be a spectrum, an image, a time series, or a higher dimensional combination of those. This model attempts to identify the different aspects that fully describe either a single observation of the sky, or a dataset derived from a number of observations. It therefore represents a description of all the metadata that may be required by both data discovery and retrieval services and data analysis applications. Metadata in this document means any data associated with the observation except for the astronomical measurements themselves.

An observation can be broken down into three main categories – Observation Data, Characterization and Provenance. Observation Data describes the units and dimension of the data. It inherits from the Quantity data model which assigns the units and metadata to either single or arrays of values. Characterization describes how the data can be used. It can be broken down into Coverage (within what limits the data is valid) and Resolution and Precision (different aspects of how accurately we are able to measure any single value). Provenance describes how the data was generated. This includes the telescope/instrument configurations, calibrations, the data reduction pipelines and the target itself.

The Observation DM can be used in different ways depending on the context. In frame of the DAL (IVOA Data Access Layer), the DM will provide standard tags to formulate a query to a VO-compliant data provider (the Coverage part of the model described below will play a frequent role here) and a standard to describe the results of such a query (like the metadata tree used in IDHA). In the context of data processing and analysis, the DM will provide a standard way to describe the accuracy, the resolution and the sampling applied to any observation. This lets tools handle observations from different archives in a systematic way. The description of the instrument configuration used to collect the data is useful in a variety of analysis and query contexts.

8.2.4 Simple Spectral Data Model

This is a data model describing the structure of spectrophotometric datasets with spectral and temporal coordinates and associated metadata. This data model may be used to represent SED (spectral energy distributions), spectra, and time series data. Spectra are stored in many different ways within the astronomical community. The IVOA model presents an abstraction for spectral data. It is required to represent a single 1-dimensional spectrum, time series photometry, spectral energy distributions which consist of multiple spectra and photometry points.

Spectral data model is based on such concepts as Spectrum and Time Series, Spectral coordinate, Flux (Spectral Intensity) Object, BackgroundModel Object, Time coordinate, Position coordinate, Accuracy Fields. Associated Metadata Fields include Coverage Fields, Frame fields, Derived Data Fields, Curation model, Data Identification model.

The Spectrum model involves objects addressed by the proposed VO Observation and Quantity data models. A single Spectrum maps to the Observation model, which will include the Curation and Coverage objects. The Flux and the spectral coordinate entries together with their associated errors and quality will be special cases of the Quantity model, as will the simpler individual parameters.

FITS serialization, VOTable Serialization and Direct XML serialization are defined for the spectral model.

8.2.5 Simulation Data Model

A data model for simulation data (named 'Simulation') is being developed within the framework outlined by the Observation model. The three main sub-categories – Simulation Data, Characterization and Provenance are still applicable. However, for simulation data it is the Provenance object, rather than Characterization that contains the real descriptive content of the model.

This object remains essentially the same as in the Observation model – a subclass of the Quantity object, used to contain the main data output of the simulation. However, for simulated data there is potentially a much wider range of quantities to be stored. In Observation at least one quantity in the data must be an observable; this is not the case in Simulation. The metadata structure – the set of UCD's used to describe each quantity must be enlarged to incorporate data clearly labelled as being 'theoretically derived'.

The Provenance object contains most of the information describing the simulation. This is because, unlike during an observation, most of the effort in acquiring the data is not through measurement but through the execution of numerical routines, thus creating the data set. The Provenance object is defined as 'the description of how the dataset was created' which for a simulation is possible to describe entirely.

Provenance can be broken down into the Theory, Computation and Parameters. Theory describes the underlying fundamental physics upon which the simulation is based. Computation describes the technique used to evaluate the physics described in Theory through the execution of numeric routines. Parameters not only define the physical context of the simulation, but also the resolution and detail. If the algorithms are analogous to a mathematical function, the parameters are the values of the input variables.

8.3 Unified Content Descriptors (UCD)

The Unified Content Descriptor (UCD) is a formal vocabulary for astronomical data that is controlled by IVOA (http://www.ivoa.net/twiki/bin/view/IVOA/IvoaUCD). The major goal of UCD is to ensure interoperability between heterogeneous datasets. The use of a controlled vocabulary will hopefully allow an homogeneous, non-ambiguous description of concepts that will be shared between people and computers in the IVO. A UCD is a string which contains textual tokens that are called *words*, which are separated by semicolons. A word may be composed of several *atoms*, separated by period characters. The order of these atoms induces a hierarchy. The UCD system is an attempt to describe simply the most commonly used quantities that astronomers want to exchange. It gives *standard names to properties of instances of concepts.*

The UCD1 set has been built at CDS Strasbourg as a collection of metadata (properties and concepts) that can be used to describe the nature of the

large astronomical data holdings at CDS. While many column headings fall naturally into the property/concept semantics, many do not. The concept description is most relevant when different concept types are involved in a context.

Our guideline in the definition of the UCD vocabulary was to study the quantities used in practice (in data tables, in FITS keywords, in the metadata of archives), and try to identify common properties and common concepts.

For example, the temperature is a property that can be measured for a telescope, the atmosphere, or for a star. We therefore define the property phys.temperature, and concepts instr.telescope, obs.air, src.star.

UCD will be used in practice for *exchanging* information using a controlled vocabulary. They are used in the VOTable standard to attach a standard description to table column names. What is needed for interoperation with other systems is a "*translation layer*" that is able to associate UCD to the parameters that are used internally, so that the output of the service contains a standard description that can be interpreted by other VO services.

Several web services have been implemented to aid in the exploitation of UCD:

- Resolver service – for given a UCD to provide a textual description of what it means;
- Listing and Browsing services allow a dynamic view of the tree of UCD;
- Search Engine allows the input of natural language, or a file of keywords, data types, and other information and tries to find suitable UCDs.

In the next version of UCD, it is planned to use it as part of a larger effort to build a semantic grid of astronomical data. This will be a large new project tentatively called UCD3. The idea is to build a semantic net that connects parameters, UCDs, names of table attributes (in multiple tables), identifiers of datasets in the VO registry, abstract grouping concepts, and so on. It is planned to try the language of the semantic web (RDF) to express relationships, and topic maps or ontology to build, expose, and reason from this knowledge.

Introducing of new UCDs on a special request can be granted by the Board controlling UCDs.

8.4 Metadata Registries for VO

8.4.1 Resource Metadata for the Virtual Observatory

A *registry* is a query service for which the response is a structured description of resources. Resource metadata constitute a "yellow pages" of astronomical information. Metadata about resources and services in VO are

standardized. *Resource metadata* are generic, high-level, and independent of any specific service. Resource metadata include:

- *Identity metadata,* which gives the resource a name and an identifier;
- *Curation metadata*, which describe who supports the resource and its availability (i.e., version, release date);
- *Content metadata*, which describe what kind of information is available (types of data, sky coverage, spectral coverage, etc.). Content metadata can be either general, applying to all resources, or associated more specifically with data collections and the services that deliver data from them.

Resource metadata are typically not queryable parameters in the underlying services, but rather they encompass information that now is simply "known" to users, or must be discovered through other means. *Service metadata* are an extension of the general resource metadata describing *how* to access the resource. Resource metadata are collected through resource registration services. The most general resource metadata is similar in concept to the Dublin Core metadata definitions (http://dublincore.org/documents/dces/).

IVOA document describes the *concepts* needed in the resource metadata. These concepts may be instantiated in a variety of standard forms, e.g. XML, UCD tags, or FITS keywords, and with a variety of mechanisms, such as Topic Maps, OWL, or RDBMSs.

Resource metadata concepts include:

- *Identity metadata* (Title, ShortName, Identifier);
- *Curation metadata* (Publisher, PublisherID, Creator, Contributor, Date, Version, ReferenceURL, Contact.Name, Contact.Email);
- *General content metadata* (Subject, Description, Source, Type, ContentLevel, Relationship, RelationshipID);
- *Collection and service content metadata* (Facility, Instrument, Coverage, Coverage.RegionOfRegard, Coverage.Spectral, Coverage.Spectral.Bandpass, Coverage.Spectral.MinimumWavelength, Coverage.Spectral.MaximumWavelength, Coverage.Temporal.StartTime, Coverage.Temporal.StopTime, Coverage.Depth, Coverage.ObjectDensity, Coverage.ObjectCount, Coverage.SkyFraction, Resolution.Spatial, Resolution.Spectral, Resolution.Temporal, UCD, Format Rights);
- *Data quality assessment* (DataQuality, Uncertainty.Photometric, Uncertainty.Spatial, Uncertainty.Spectral, Uncertainty.Temporal);

Service metadata concepts include:

- I*nterface metadata* (Service.InterfaceURL, Service.BaseURL, Service.HTTPResultsMIMEType);
- *Capabilities metadata* (Service.StandardID, Service.StandardURL, Service.MaxSearchRadius, Service.MaxReturnRecords).

8.4.2 IVOA Metadata Registry Interface

IVOA has developed the standard interfaces (http://www.ivoa.net/twiki/bin/view/IVOA/IvoaResReg) that enable interoperable registries. These interfaces are based in large part on a Web Service definition in the form of a WSDL document. Through these interfaces, registry builders have a common way of sharing resource descriptions with users, applications, and other registries. Client applications can be built according to this specification and be able to discover and retrieve descriptions from any compliant registry.

A searchable registry is one that allows users and client applications to search for resource records using selection criteria against the metadata contained in the records. A searchable registry gathers its descriptions from across the network through a process called *harvesting*. A publishing registry is one that simply exposes its resource descriptions to the VO environment in a way that allows those descriptions to be harvested. A full registry is one that attempts to contain records of all resources known to the VO. A local registry, on the other hand, contains only a subset of known resources.

The IVOA Registry Interface consists of three query operations:

- Search searches the Registry in order to obtain the VO resources.
- KeywordSearch is a helper query based on a set of key words.
- GetRegistries is another helper query to obtain Registry VO resources.

and six harvesting operations, which support resource harvesting in accordance with the OAI-PMH definition [LOAIF].

Search (ADQL) operation implements the IVOA VO resource retrieval between registries. The Search operation has only one parameter, i.e. the query. The result of the Search operation is a set of the Resource metadata wrapped in a VOResources element (a general-purpose root element encoded in XML). A VOResources element may contain a set of many Resource elements. For the purposes of Resource retrieval ADQL expression focuses only on the WHERE clause of the ADQL, and uses Xpath expressions for the purpose of querying.

The KeywordSearch (String words, boolean orValue) interface is used for a keyword based query that automatically searches against the supported registry names. All searchable registries support the required elements given in a Resource schema (Identifier, Description, Title, ResourceType, Subject, Type**).**

VOResource, in general, refers to a family of schemas that includes the core schema (VOResource-v0.10) and a set of standard extensions. These schemas have been released from the official IVOA schema distribution area. VOResource extensions are defined for Cone Search services, Simple Image Access services, for Data and Services, for Registries, for Communities (organizations and projects).

8.5 VOTable Format Definition

The VOTable format (http://www.ivoa.net/twiki/bin/view /IVOA/IvoaVOTable) is an XML standard for the interchange of data represented as a set of tables. A table is an unordered set of rows, each of a uniform format, as specified in the table metadata. Each row in a table is a sequence of table cells, and each of these contains either a primitive data type, or an array of such primitives. VOTable has built-in features for big-data and Grid computing. It allows metadata and data to be stored separately, with the remote data linked. Due to that it is possible to send metadata-rich pointers to data tables in place of the tables themselves. The remote data is referenced with the URL syntax. The data part in a VOTable may be represented using one of three different formats: TABLEDATA, FITS and BINARY. TABLEDATA is a pure XML format so that small tables can be easily handled in their entirety by XML tools. VOTable can be used either to encapsulate FITS file, or to re-encode the metadata; unfortunately it is difficult to stream FITS. The BINARY format is supported for efficiency and ease of programming.

The data model of VOTable can be expressed as:

```
VOTable = hierarchy of Metadata + associated TableData,
arranged as a set of Tables
Metadata = Parameters + Infos + Descriptions + Links +
Fields + Groups
Table = list of Fields + TableData
TableData = stream of Rows
Row = list of Cells
Cell = Primitive or variable-length list of Primitives or
multidimensional array of Primitives
Primitive = integer, character, float, floatComplex, etc.
```

A table cell can contain an array of a given primitive type. The overall VOTable document structure is described and controlled by its XML Schema referenced at its top. Basically, a VOTable document consists of a single all-containing element called VOTABLE, which contains descriptive elements and global definitions (DESCRIPTION, COOSYS, PARAM, INFO), followed by one or more RESOURCE elements. Each Resource element contains one or more TABLE elements, and possibly other RESOURCE elements.

8.6 Data Access Layer

8.6.1 DAL Architecture

The task of the IVOA DAL working group is to define and formulate standards for uniform access to VO data that may have heterogeneous representations by different data providers. Architecturally DAL (http://www.ivoa.net/twiki/bin/view/IVOA/IvoaDAL) consists of a family of data access services that provide access to both data and computation – access to VO resources. Client data analysis software will use these services to access data via the VO framework; data providers will implement these services to publish data to the VO. Principal data types within the scope of the DAL and mapping of data types to access services (e.g., image, table, spectrum, time series, etc.) are to be defined. Each service supports a specific data model and is implemented through the respective data access protocol. Access protocols form a middleware between the VO resources and client data analysis programs. Thus distributed multiwavelength data access and analysis is planned to be developed.

Current DAL services include:

- Cone search;
  - access to astronomical catalogs;
  - simple query based on position, search radius;
  - returns a VOTable containing results;
- Simple Image Access (SIA);
  - uniform access to image archives;
  - atlas and pointed image archives;
  - image cutouts, image mosaics;
  - image is returned as a FITS file or graphics file;
- Simple Spectral Access (SSA, currently being specified);
  - access to 1D spectra and SEDs;
  - spectra is returned as ASCII, VOTable, FITS.

Basic DAL Service Interface includes:

- Registry used to discover services or data;
  - registry query is optional;
  - data query will locate services providing desired data;
  - service query will find services by type;
- Data query against specific service;

    - returns a VOTable listing datasets which satisfy query;
- Dataset access;
    - returns an individual object of the specified type;
    - service metadata query;
    - describes the capabilities of an individual service.

DAL might be considered as a step to the mediator layer considered for RVO. Main difference is that DAL is oriented on support of limited number of built-in data types by the respective wrappers (adapters) implemented as services. This makes DAL quite easily implementable. RVO is more concerned with user-defined, arbitrary mediator types that can be queried, with wrappers located at the bottom level of the mediators.

8.6.2 Simple Image Access Protocol Specification

This specification defines a protocol for retrieving image data from a variety of astronomical image repositories through a uniform interface. A query defining a rectangular region on the sky is used to query for candidate images. The service returns a list of candidate images formatted as a VOTable. For each candidate image an access reference URL may be used to retrieve the image. Images may be returned in a variety of formats including FITS and various graphics formats. Referenced images are often computed on the fly, e.g., as cutouts from larger images. Data collections are often distributed, and the client may query multiple image services simultaneously, e.g., to gather data from multiple wavelength regimes or surveys to analyze a single region on the sky. The client describes the ideal image – what it would like to get back from the image service – and the image service returns a list, encoded as a VOTable, of the (often virtual) images it can actually return. A key point is that it is entirely up to the image service and what images, if any, it offers to the client. These images may range from a simple list of static archive images which intersect the region of interest defined by the client, to a mosaiced and reprojected synthetic image matching the ideal image requested by the client.

This specification is based primarily on two documents. The first document, "Simple Image Retrieval: Interface Concepts and Issues", describes a longer term view of how simple image access can fit into a more general framework for image access in the VO. The URL-based implementation specified here is intended to be consistent with the concepts discussed in this document. A prototype SOAP/WSDL based Web Services implementation is also planned. The second document, the "Simple Cone Search specification" provides a means to query catalogs via HTTP with a uniform interface. The Simple Image Access interface (SIA) defined here follows a similar to Cone Search approach.

The image data model assumed is minimal at this point. An image should be a calibrated object frame imaging some region of the sky. Only two

dimensional images are fully supported within the interface at this time. Images can be returned as either FITS files or as graphics images. Ultimately, VO data models will provide a means to describe more complex data objects within the VO than be directly addressed by the SIM prototype.

Compliance with this specification requires that an image access web service be maintained with the following characteristics:

- The Image Query web method searches for available images that match certain client-specified criteria. The response is a table that describes the available images, including image metadata and access references (implemented here as URLs) for retrieving them.
- The Image Staging web method (optional) allows clients to direct the image service to stage selected images for later retrieval.
- The Image Retrieval (getImage) web method allows clients to retrieve single images.
- The image service MUST be registered allowing clients to use a central registry service to locate compliant image services and select an optimal subset of services to query, based on the characteristics of each service and the image data collections it serves.

8.6.3 Simple Spectral Access Specification

The goal of the Simple Spectral Access (SSA) specification is to define a uniform interface to spectral data including spectral energy distributions (SEDs), 1D spectra, and time series data. In contrast to 2D images, spectra are stored in a wide variety of formats and there is no widely used standard in astronomy for representing spectral data.

The data model for spectral energy distributions defines a set of spectra or time series, some of which may have only one or few data points (photometry) and each of which may have different contextual metadata like aperture, position, etc. A *SED* object has a number of global attributes indicating the number of SED segments and *curation* information. Each *segment* has a *frame*, *coverage*, *curation* and *data identifier* object. The *frame* object is a simplified instance of the space-time coordinate system object The *coverage* object holds info about the observed region on the sky, the time range and spectral range. The *time coordinate* contains elapsed times relative to a reference time. The *spectral coordinate* can be expressed as a wavelength, frequency or energy plus velocity.

The purpose of a spectrum query is to determine the availability and characterization of data satisfying the constraints. The result is encoded as a VOTable. Queries can be restricted to certain types of data using the keywords *findSED*, *findSpectrum*, *findTimeSeries*. Technically based on SOAP/HTTP, an SQL query is generated. The format of the data returned in the retrieval mode

could be a VOTable, FITS, native XML, a graphic file or some *foreign* format used by a data provider.

Compliant SSA services should be registered with a VO registry. A registry stores service metadata which characterize it and any associated data collections. Capabilities should also be described. The gathering of service and capability metadata from all such services enables a client to use the registry to discover the services most appropriate for a particular computation.

8.7 IVOA Query Language

8.7.1 IVOA SkyNode Interface

The SkyNode Interface describes the minimum required interface to participate in the IVOA as a queryable VONode as well as requirements to be a Full OpenSkyNode, part of the OpenSkyQuery Portal. OpenSkyQuery (http://www.openskyquery.net/) opens up the SkyQuery protocol to enable other databases and servers to become "Full SkyNodes". It should be noted that the SkyNode Interface is also related to Data Access Layer WG of the IVOA.

The Astronomical Data Query Language (ADQL) is considered as an XML document format for transported queries to IVOA SkyNodes. Different SkyNodes may not support all features of the Language. Hence ADQL would be passed from the SkyQuery Portal to the SkyNodes or it may come directly from a client or the VOQL portal. All nodes and the portals should be accessible via SOAP services. Additionally for the Open SkyQuery Portal some form of string based query like the current SkyQL (www.skyquery.net) would be accepted. A parser would easily convert this to ADQL, i.e. SkyQL would have the same semantics as ADQL but the syntax would be an SQL like string rather than XML.

Basic SkyNode (see the Table) is the minimum IVOA SkyNode Interface – this is useful in itself as it allows one to send queries to a system using ADQL. This is also just one step up from cone search. A matrix has been used as any feature on their own may be useful, i.e. a node which can do XMATCH is already useful even if it may not participate in the portal because it lacks other features. It is assumed large surveys would be Full SkyNodes.

Feature	Basic SkyNode	Full SkyNode	Optional
ADQL – circle	X	X	
Functions	X	X	
Xmatch		X	
Performance Query		X	

Takes exec plan		X	
Footprint – Region intersect		X	
MYDB			X
Authentication – must have with MYDB			X

Basic SkyNode should provide the following capabilities:

- SkyNodes shall register with the registry with type=”SkyNode” specifying also whether UCDs are acceptable;
- SkyNodes shall implement the “Tables” interface, which returns a list of all tables that may be used in ADQL;
- SkyNodes shall implement the “Columns” interface, which returns information about the columns of a given table name;
- SkyNodes shall implement the “Formats” interface. This takes no parameters and returns a list of formats which this Node supports for Query Results (VOTable, DataSet, ASCII);
- SkyNodes shall implement the “Functions” interface;
- SkyNodes shall implement the "PerformQuery" interface, which takes an XML document including an ADQL query and an optional string parameter called “format”;
- SkyNodes should accept the VOQL equivalent of Standard SQL-92 Metadata queries.

Full SkyNode capabilities include:

- SkyNodes shall implement QueryCost() interface which takes a simple ADQL query to return the object density per square degree for a set of criteria;
- SkyNodes shall accept complex Shapes in their queries as defined in the ADQL syntax;
- SkyNodes shall be able to perform cross matching between their survey and table of data provided in VOTABLE format;
- SkyNodes shall implement an ExecutePlan() web method, which takes an ExecPlan and passes the relevant part of the plan to the next node. From that node it shall receive a VOData result. If there are no more nodes in the plan it simply executes the ADQL query and returns the resulting VOResult;

- SkyNodes should implement the footprint service. This would take a region specified in the region XML and return a new region which is the intersection of the survey an the given region.

8.7.2 Astronomical Data Query Language (ADQL)

ADQL is based on a subset of SQL plus region with, as a minimum support, for circle (Cone Search). Services for translation of SQL to ADQL and back may be found at http://skydev.pha.jhu.edu/develop/vo/adql/. ADQL is designed to be the request format of the OpenSkyQuery protocol. The OpenSkyQuery protocol drives a data service that allows querying of a relational database or a federation of databases. In this case, the request is written in a specific XML representation of ADQL.

ADQL has two forms:

- *ADQL/x*: An XML document conforming to the XSD;
- *ADQL/s*: A String form based on SQL92 (the BNF exactly defines the form of SQL92) and conforming to the ADQL grammar. Some non standard extensions are added to support distributed astronomical queries.

The XML expression of ADQL (ADQL/x) is recommended in the Virtual Observatories for communications between portals and data servers. The string version of ADQL (ADQL/s) is more suitable for human to understand the queries. Extensions to SQL92 include:

- ADQL supports the region specification as defined by the region.xsd in Space Time Coordinates for VO (http://www.ivoa.net /internal/IVOA/InterOpMay2003DataModel/STCdoc.pdf and http://hea-www.harvard.edu/~arots/nvometa/STC_UML.pdf). The Region would look something like: *Region('CIRCLE J2000 19.5 – 36.7 0.02')*. Other constructs (see [SQLHTM]), such as RECT, POLY, and CHULL, are also supported;
- JDBC Mathematical functions shall be allowed in ADQL as follows: Trigonometric functions: acos, asin, atan, atan2, cos, cot, sin, tan; Math functions: abs, ceiling, degrees, exp, floor, log, log10, mod, pi, power, radians, sqrt, rand, round, truncate;
- XMATCH implies crossmatch between two or more astronomical catalogues. XMATCH appears in the Where clause and looks like a function. Each parameter is a table to be crossmatched, the final parameter is the sigma value for the chi-square match;
- To support Xquery as well as SQL, and since some of our data formats are described as XSD, it will be possible to express selections and selection criteria as a simple Xpath. Square brackets([,]) and standard operators such as parent are NOT supported;

- ADQL supports the top syntax to return only the first N records from a query;
- ADQL allows units for all constant values specified in the query. These are optional.

Sample applications and tutorials for development and deployment of ADQL services are available at http://skyservice.pha.jhu.edu/develop/vo/adql/.

ADQL servers will be integrated into easy-to-use portals. Typically a Portal queries the registry to find SkyNodes, then interrogates their relational database with the OpenSkyQuery protocol. A client can first request the table names and descriptions, then request the schema of any of the tables, then build a suitable query to select data. ADQL also allows joins across SkyNodes. Similar actions can be done by preparing, e.g., a GUI that creates equivalent ADQL scripts. The Portal will formulate a plan and create multiple queries, typically one per archive. And the results are collected, joined, and served to the users. ADQL will eventually be integrated into all the data services to provide an advanced query capability. Both simple parameter-based queries and query language-based (ADQL) queries will be provided.

8.7.3 VO Query Language

The Virtual Observatory Query Language (VOQL) is an ambitious language at a higher level than ADQL (http://www.ivoa.net/twiki/bin/view /IVOA/IvoaVOQL). A VOQL portal would take VOQL programs. This would need all the work of the SkyQuery portal and more to make it function. In summary we may see 3 layers of VOQL:

- VOQL1 WebServices : ADQL and VOTABLE to exchange information between machines;
- VOQL2 Federation : SQL-like query language and federation system, i.e. combination of SkyQuery , JVOQL and VO standards;
- VOQL3 SkyXQuery: future XML-based query language.

The highest level of VOQL is a semantics-based language that allows astronomers to build queries in the language of astronomy rather than the language of databases. Efforts with an ontology of units allows queries expressed in one unit to engage resources expressed in another unit. Similarly astronomical coordinates can be fungible, so that a query in equatorial coordinates can return a resource expressed in galactic coordinates – but in the correct part of the sky. A similar approach allows federation of spectral data that uses different spectral coordinates.

This level of semantics, describing the structure of astronomical datasets, interacts with the astronomical semantics provided by the UCD schema to quantify use of astronomical knowledge. For example, a data model to define spectra may specify that a spectrum has an array of data representing an

observable quantity and an array of values representing the spectral coordinate. The UCDs associated with an instance of this data model will specify whether that particular spectrum has an observable of flux or surface brightness, and a spectral coordinate of frequency or wavelength. A data model may also represent a higher level resource such as a compute service, in which the input parameters required by a particular class of service such as source detection programs are defined. Again, the values of some data model metadata may be UCDs which describe what kind of parameters are to be returned by the source detection.

9 Requirements for the RVO infrastructural components oriented on RVO usage in education

Virtual observatories are considered as *collective memories* (CM) converging sources of various kinds (section 4). To *publish an information product* means to make it available in VO through services that are accessible via VO portals. Facilities for publishing of the results of observations and simulation have been analyzed so far. Discovery and manipulation are basic kinds of operations applicable to such products. Database technology is mainly used to support them. For publishing the results of research as well as for education purposes another technologies are required that are based on the concepts of Digital libraries. Digital libraries are organized collections of information resources in digital or electronic format along with the services designed to help users identify and use those collections. Digital libraries promise to provide effective information services by offering the following advantages: faster delivery, a wider audience, greater availability, more timely information, more comprehensive. Specific class of digital libraries is formed by Digital Libraries in Education (DLE) [KDLEU]. It is suggested to use DLE approaches for the RVO infrastructures intended for RVO usage in education.

A DLE is envisaged as a *comprehensive library of the digital resources and services* that are available for education in science, mathematics, engineering, technology, and other disciplines. The key word here is "comprehensive." Faculty are very specific in wanting a single place where they and their students can discover, use, and possibly contribute a wide range of materials. A DLE is considered to be a *federation of library services and collections* that function together to create a digital learning community. The following long-range objectives for DLEs are usually formulated:

- Life-long learning;
- Learning anytime anywhere;
- Increasing the quantity, quality, and comprehensiveness of Internet-based science educational resources;

- Making these resources easy to discover and retrieve for students, parents, and teachers;
- Ensuring that these resources are available over time.

DLEs take a broad view of science and technology, and of scientific education. The primary audience is faculty and undergraduate students, but there is no hard distinction between the needs of high school students, undergraduates, and graduate students, nor between students in formal programs, independent learners, and the general public.

DLEs also open the opportunity for students at different institutions to work on *joint projects or experiments*, perhaps sharing and adding to the same data set and its analysis. This would also promote *physical resource sharing*, as students and instructors may have varying access to high-end instrumentation, computational capabilities, data collections, and technology.

DLEs should provide various services, such as cataloguing, archiving, selective dissemination of courseware and other instructional materials developed internationally, annotation, evaluation, cross-lingual search and retrieval, personalization, recommendation, instructor support, and copyright management.

DLEs provide facilities for long-term, distributed, and stable repositories for data and metadata that institutionalize public-domain data holdings. These repositories must provide quality control, interchange formatting, and translation, as well as tools for data preparation, fusion, mining and knowledge discovery, and visualization. A key element associated with filling this need is the development of middleware and related data storage strategies.

VO provides an exceptional opportunity to link observational and simulation sources collected by specialized organizations on the national level or globally with publications and education-oriented sources. An access to expensive scientific instruments (e.g., telescopes) and to specific service can also be provided. General VO infrastructure (Grid based) provides the required techniques for federated access to massive data collections. Such possibilities create efficient conditions to involve learners in research in the early stages of their education.

In VO large sky surveys are based on catalogs and portals that provide most of the functionality of digital libraries (collection organization, search, access, and manipulation). For researchers, additional technology is needed to enable small collections, in which images and results can be published for use by the broader community. The approaches that are being tried within the NVO are based on technology from the grid community and the digital library community.

The Globus project at ISI is developing a Metadata Catalog Service for registering metadata about user files. The system will provide interfaces based on the OGSA-DAI framework (http://www.ogsadai.org.uk). Details about the

current version of MCS are found at http://www.isi.edu/~deelman/MCS. The data grid project at SDSC uses the SRB (Storage Resource Broker) technology to support distributed collections. The access interfaces include Open Archives Initiative metadata harvesting protocol, WSDL services for metadata and data management, and web browser interfaces to support search. Both projects are designed to scale to millions of digital entities. The SRB technology supports federation of independent data grids, making it possible for a major sky survey to implement their own data grid, and then at a later time when the data is publicly available, publish their data grid into the NVO data grid (http://www.npaci.edu/DICE/SRB).

A second approach uses the standards and open sources implemented within the digital library community. The DSpace technology from MIT (http://www.dspace.org/) focuses on preservation, and the Fedora technology from Cornell (http://www.fedora.info/) focuses on manipulation. Both approaches are being evaluated within the NVO. JHU is collaborating with Fedora to build mySpace for storage of small collection. SDSC is collaborating with MIT to integrate the DSpace technology as an access mechanism on top of the SRB data grid. Both projects will provide web-based interfaces for accessing collections that are deposited by researchers. The DSpace/SRB integration will also provide the ability to federate multiple independent digital libraries, making it possible for an institution or researcher to establish their own collection, and then federate that collection with a larger publication environment. These point towards two possible approaches to support for mySpace collections:

- NVO central repository. This would be similar to the NSF National Science Digital Library, with a single repository pointing to all material, backed up by a persistent archive for preservation.
- NVO federation. This is similar to the approach being followed by the Worldwide Universities Network for the sharing of research data between institutions. The WUN serves as the policy management system for federating multiple independent digital libraries. NVO can provide a similar role, serving as the institution that defines and manages the federation policies required for publication of researcher data into the NVO name space.

Equivalent metadata catalog services are incorporated in sky survey catalogs, the SRB data grid, NVO portals, and the SRB federation environment. At some point, NVO will need to do a comparison between the approaches to analyze which system provides the best support for astronomy collections. The comparison needs to address federation of name spaces between independent data grids, support for preservation environments, and support for NVO portals.

The NVO experience is important for extending basic VO technology with publishing of small, sometimes individual collections. For the basic DLEs functionality the National Science Digital Library (NSDL) project as well as

the Digital Libraries for Earth Science Education (DLESE) [DLESN] can serve as sources of technological and organizational ideas.

Similarly, a Digital Library for Astronomy Education (DLAE) might be an adequate solution. Its mission could be the improvement of the quality, quantity, and efficiency of teaching and learning about the Universe, by developing, managing, and providing access to high-quality educational resources combined with the VO resources and supporting services through a community-based, distributed digital library. DLAE should provide a mixture of resource discovery using keywords, grade level, educational resource type descriptors (atlas, visualization, activity, lab, etc.), as well as spatial and temporal footprints and astronomy-referenced information (similarly to gazetteer services in GIS environment). More advanced curriculum-based resource discovery is required to allow for location of resources related to certain educational topic defined in natural language or conceptual graph. The DLAE catalogue management system and Open Archives Initiative (OAI) facilities should be based on the respective open sources (e.g., DSpace or Fedora).

Metadata standards for DLAE can be developed taking into account NASA, NSDL, DLESE, ADL (the US Department of Defense's Advanced Distributed Learning Network) experience. Learning object metadata standard [DLOMN] should also be taken into account. This standard specifies a conceptual data schema that defines the structure of a metadata instance for a learning object. For this standard, a learning object is defined as any entity, digital or non digital, that may be used for learning, education, or training. For this standard, a metadata instance for a learning object describes relevant characteristics of the learning object to which it applies.

Other experience, such as VTIE (for Virtual Telescopes in Education) project, is to be analyzed. VTIE aims to bring observational astronomy directly to learners in both formal and informal settings by providing tools for both educators and students. For educators, VTIE provides the capability to design astronomy experiments, an online review tool to comment upon students proposals and papers, and classroom management tools (e.g. messaging service and ability to create a reading list). For students, VTIE provides an interface for developing an observing proposal (details of which are designed by the educators), access to online data services, an online observing log, and a Paper Writing Tool to complete the process by reporting their results.

Close to this project is an idea of creation of an Open Virtual Media of astronomical education to transit from separate astronomical scientific – educational schools existing at universities, to common educational space integrating an intellectual potential, methodical and research possibilities of educational and scientific organizations.

Attempt of realization of such approach is developed at SAO's Shared Center of Science and Education as the project of creation of an Open Virtual

Media of astronomical education. SCSE SAO of RAS was created in 1997 on the base of Special Astrophysical Observatory of Russian Academy of Science. This center unites efforts of the five largest universities of Russia: Moscow, St.-Petersburg, Rostov, Kazan and Ural in the training of the students and post-graduate students of astronomy and preparation of highly qualified astronomers. The center provides an interaction between the academic science and educational process in higher educational institutes.

Besides technological issues, specific organizational measures are required to provide for efficient and sustainable development of DLAE [KDLEU]. For instance, important resources for research and education include not only archives of data but also streams of simulation data, as well as software services. Such resources should become part of DLAE content. A special center is required that would offer software and services that enable universities to acquire and use astronomic and related data on their own computers, often in real time or "near-real time"—that is, the data might be sent to participants almost as soon as the observations are made. Such center software and services are available to any university at no cost. Member institutions provide their own computers, network connections, human resources, and other requirements for participation, including access fees for certain data. Through computer networking, this Center participants become members of a mutually supportive "virtual community"—a nationwide group of electronically linked individuals who hold common academic interests in the astronomy and related sciences and who share similar needs for data and software. Similar infrastructure is provided by the Unidata center for atmospheric research [KDLEU]. Unidata was founded in the atmospheric science ADL experience, learning object metadata standard [DLOMN] domain; however many universities employ Unidata systems or data to support education and research that falls outside that domain. This reflects a trend toward interdisciplinary education and research.

To provide a competitive education in natural sciences (specifically, in astronomy), different countries may establish their own DLEs (e.g., as a national DLE, collaboratively with other DLEs, or as a regional DLE). They cannot passively wait until suitable global digital educational content is formed. The digital content of DLEs remains dependent on the language (or language groups) used by the educational community in each country, as well as the culture and national traditions in education. A significant amount of time is required to form the national community around DLEs, collect the DLE content, and educate specialists to develop, maintain, and govern DLE. These considerations are to be taken into account for the RVO planning. In particular, DLE is distinguished from other ICTs applicable to education (e.g., multimedia, distance learning) by several important features:

- Metadata for the DLE should be consolidated;
- To establish a DLE (after technology is installed as software components), serious efforts are required to collect (harvest, integrate,

gather, register) the digital resources, and to maintain and continuously extend them. If the digital content is not completely borrowed from another DLE, this process requires specific organizational efforts and investments. In particular, for astronomy enormous amount of resources exist around the world that deserve of being included into the DLAE. At the same time, it is not a task that can be done by a separate individual (the way an educator can individually establish and use multimedia technology preparing courses). Governance, maintenance, and a community must be arranged around the DLE to make it sustainable;

- To make a DLE useful, additional efforts are required to provide for preserving the proper quality of the digital content. This is also not an individual effort. Various organizations in society must be involved in the process of creating digital content of the required quality;
- To make information in digital form widely available requires supporting rights of access and use, including copyright, preservation of the integrity of the document, licensing, and payment for use;
- In DLEs with digital content a wide set of interrelated services require administration and development.

10 Subject mediation infrastructure as a basis for problem domains representation in RVO

10.1 RVO Subject Mediation Concepts and Facilities

10.1.1 Principles of subject mediation

The current and projected rate of data volume growth from measurements and observations in astronomy is exponential. The widening gap between the scientists and the sources of the data requires a major paradigm shift in the way of scientific problems solving over multiple large distributed data sources and services (that are concentrated in specialized centers of data and computational facilities). Various technical infrastructures are intended for this way (such as Web-services, Grid architectures, middleware frameworks). In spite of that, the problem of efficient integrated representation of multiple sources for a researcher remains to be open. Two principally different approaches to this problem exist: moving from multiple sources to a researcher and problems (in this case an integrated representation of multiple sources is created independently on the tasks) and from a problem to sources (in this case a subject domain definition for a class of problems is created, the relevant sources are identified and mapped into this definition). The first approach is not scalable. For example, currently in astronomy the number of existing sources (catalogs, archives) has an order of many thousands. The integrated schema of

a set of sources is to be modified each time when a new source appears. For such multi-database (to each source its subschema in the global schema corresponds, as it is done in SkyQuery) the global schema becomes hardly embraceable by a researcher. Currently this first approach to information sources integration dominates in IVOA.

Another approach assumes creation of mediators supporting interaction between a researcher and relevant data sources and services through a subject domain description for a class of problems [KSMDL]. Scientists spent centuries to develop efficient instruments, well-defined concepts, theories and computational models in various branches of science. A subject domain definition for a class of problems roughly is subdivided into the three interrelated parts: concept spaces, theories, models, hypotheses; results of experiments, observations, measurements, simulations; computational models and tools. Publications, curricula and educational modules are added to each part of this framework. Definition of a subject domain in terms usual for researchers should serve as a base (an abstraction) for the mediator specification independent of existing data sources and services. *Subject mediators* are emphasized that support representation and access to various subject domains in astronomy. They lead to more knowledge-based organization.

Based on such abstractions, an intermediary layer is formed by subject mediators providing metainformation uniformly characterizing their subject content. A *canonical information model* is used for the mediator definitions making possible to query such abstract content and compute the result. Introducing of such mediators frees the researchers from having to identify relevant entities among multiple heterogeneous sources and services defined in non-uniform terms, semantically reconcile and correlate them, formulate research tasks in unusual terms, structures, functions and processes. Each mediator supports the process of systematic discovery and registration of sources uniformly expressing their definitions in terms of the mediator. This process of registration is assumed to be semi-automatic. The mediator's canonical information model should be powerful enough for uniform equivalent and complete representation of various data sources and services.

Two separate phases of the mediator's life cycle are distinguished: consolidation phase and operational phase. *Consolidation phase* is intended for creating a definition of the mediator. During this phase a consensus in the subject community should be reached on the mediator ontology, concepts, data structuring and behaviors to consolidate definitions of the subject theories and models, observable (measurable) characteristics of real world objects, description of methods and instruments for observation, measurement and experimental data, simulations, data analysis results, problem definitions and methods of solution, algorithms and programs. Integration of such information is driven by scientific and educational needs. This process is completely

independent on pre-existing information sources. During the *operational phase* the sources relevant to a subject mediator are dynamically registered in it and tasks can be executed. During registration, the source capabilities (ontologies, types, functions, query language capabilities, etc.) are expressed in terms of the subject mediator's metainformation and respective wrappers are developed (if required).

Mediators form a recursive structure: a mediator can be considered as an information source for other mediators. Architecturally, a mediator is a service in the grid.

10.1.2 Mediation methods

Mediator supporting procedures include:

1. Discovery and registration in mediator of relevant information sources (data sources and/or services)

A source registration [BREGM] is a process of purposeful transformation of specifications including decomposition of the mediator type specifications into consistent fragments; search among specifications of sources of relevant data types – candidates for *refinement*[3] by them of the respective mediator types; composition of rules, defining source classes as compositions of the mediator classes. Value, structural and behavioral conflicts reconciliation should also be provided.

Discovery of relevant data types in sources is based on three models: metadata model describing the information sources, ontological model defining concepts of a subject domain and canonical model providing for definition of structure and behavior of objects of the subject domain and of the source. Decisions in canonical and ontological models are based on semantics of canonical model and on a proof of refinement relationship between type specifications. Decisions in the metadata model are based on non-functional requirements for the required sources (in particular, data quality is one of them).

A prototype for heterogeneous information source registration at subject mediators has been developed [BREGM]. Source specifications as materialized views over virtual classes of mediator are designed applying compositional development method (source definition is treated as a specification of requirements and class definitions of the mediator schema are treated as component specifications). This approach is intended to cope with a dynamic, possibly incomplete set of sources. Sources may change their exported schemas, become unavailable from time to time. To disseminate the

[3] Refinement of type *A* by type *B* means that user can apply type *B* instead of type *A* without noticing difference resulting of such substitution.

information sources, their providers should register them at a respective subject mediator. Such registration can be done concurrently and at any time. The method is applicable to wide class of source specification models. Ontological specifications are used for identification of mediator classes semantically relevant to a source class. Maximal subset of source information relevant to the mediator classes is identified (due to identification of maximal commonality between a source and mediator level class specification). Concretizing types are defined so that the mediator classes instance types are refined by the source instance type. This direction naturally supports query rewriting refining a mediator query in terms of a specific source. Such inversion is natural for the registration process: a materialized view (requirements) is constructed over virtual specifications (components).

2. Problems formulation and execution over the multiple information sources

The approach consists in problem formulation in terms of subject domain specification and rewriting of this formulation into a plan – an ordered set of subtasks and queries to the relevant information sources registered at the mediator. For problem formulation recursive queries can be applied. A method for such query rewriting in the typed object environment is required. The approach is based on the refinement relationship between data types of mediator and sources. Thus a containment of rewritten rules in the original ones can be reached. The execution plan is to be optimized so that data should not be migrated to end users but rather accessed, processed, and explored remotely across the grid in a set of distributed data sources and computational services. Such policy for doing computations near data is well agreed with data intensive calculations (e.g., in astronomy it is rare to find computations of more than 10,000 instructions per byte).

A query rewriting method for the heterogeneous information integration infrastructure formed by a subject mediator environment has been developed [KREWR]. The approach treats schemas exported by sources as materialized views over virtual classes of the mediator. Queries are considered in frame of an advanced canonical object model of the mediator. The method provides for establishing of the query containment relationship in the typed environment.

Various problems can be defined using such declarative way of their formulation. Alongside with that, it is required to provide a possibility of call of such declarative facilities from programming languages.

3. Subject mediator specification and problem formulation

Methodology (as a collection of principles, models, methods and rules) for subject mediator specification and problem formulation in the environment of multiple heterogeneous distributed information sources is required. Canonical information model of the mediator is to be used as a tool for that. Subject domain definition is assumed to be multi-layered with the layers of descriptive

models, descriptive mathematical models, formalized mathematical models, computational models. These layers range from purely verbal to purely formal descriptions. Kinds of non-functional requirements for a class of problems, ways for definition of concepts (ontologies), data structures, functions and processes are to be defined. Applicable services include mining the individual data sources looking for patterns, cross-correlating sources to find new phenomena, simulation facilities. Cross-correlation (usually under spatial and timing constraints) imposes serious requirements for data movement and computation. Procedures for reaching consolidation of a subject mediator specification in a community, for validation of this specification on the basis of representative information sources, for reaching of specification completeness and their consistency are required.

4. Portals

Methods for user interface generation using mediator metainformation are to be investigated and developed. Visual modeling of scientific artifacts and an ability to generate executable mathematical models for data simulation and analysis are key objectives for the portal development. Another requirement is that portals should provide an access to mediated data sources alongside with publications and courseware organized by means of digital libraries interoperating through OAI.

10.1.3 Subject mediation tools

Mediator level supporting tools include:

1. Tools for sources discovery and registration should include:
   - Metainformation base support (including loaders of the mediator and information source definitions);
   - Facilities for ontological contexts reconciliation;
   - Facilities for the relevant sources and their type fragments discovery;
   - Information source views generators;
   - Data model mapping and wrapper generators.
2. Tools for problems formulation and execution should include:
   - Query rewriter;
   - Planner;
   - Task program execution engine.
3. Portal supporting tools should include:
   - User interface generation tools based on the mediator metainformation and task formulation;
   - Facilities for the mediator metainformation visualization;

- Executable mathematical model generators.

10.1.4 Subject mediation approach and the IVOA architecture

IVOA conceptual modeling vs subject mediation modeling

The document "A unified domain model for astronomy" is the IVOA attempt to define a conceptual model for the whole astronomy. The model is intended to be used as:

- the basis for a meta-data repository that archives can use to describe their data products in a common model;
- a model describing the entities (classes and attributes) that can be used in a common query language for these astronomical archives and for the relations that can be followed from these entities in navigation to related ones.

First of all, note that in IVOA such sort of data models is not supported by adequate architectural idea or a technology. Due to that the model (even after its consolidation – reaching a consensus on the proposed definitions in the astronomical community) has little chances to be properly used. On the other hand, it is the subject mediator architecture that is required to support such way of modeling.

According to the subject mediator architecture, the conceptual model can be used as a mediator schema consolidated in astronomy independently on existing data sources and services. According to the mediation principles, existing sources/services should be registered at the conceptual model, mapped to it so that querying through the domain definition of the registered sources becomes possible.

Other documents prepared by the IVOA Data Modeling Group are more modest, though quite diverge w.r.t. the domains selected. Simple attempt to classify the domains follow:

- "built-in" data types, characteristic for astronomy: Spectra, Quantity, Passband, Interferometry, Transforms;
- class of observation data: Observation;
- specific class of problems: Simulations.

Independently on the domain, the models are intended for unification of the definitions, on reaching a consensus on the definitions of concepts (types) in one or another domain. For instance, the Observation data model is intended for providing a standard way to describe the accuracy, the resolution and the sampling applied to any observation.

It can be easily seen that again, these definitions can be used as a specification of a part of mediators the subjects of which include the respective domains.

It is possible to conclude that the subject mediation approach can contribute to the IVOA Data Modeling Group not only a proper vision of supporting technology, but also a methodology of proper selection of domains to be defined and better focus of what and how is to be defined for the domain. Specifically, in RVO it is intended to identify classes of research problems in astronomy and to define subject mediators (conceptual data models according to the terminology of the IVOA Data Modeling Group) for the respective domains. With such clear intention in mind, the Simulation problem class might require serious reconsideration.

DAL layer vs Mediator layer

The task of the IVOA DAL working group is to define and formulate standards for uniform access to VO data that may have heterogeneous representations by different data providers. Architecturally DAL consists of a family of data access services that provide access to both data and computation – access to VO resources. Each service supports a specific data model and is implemented through the respective data access protocol. Principal data types within the scope of the DAL and mapping of data types to access services (e.g., image, table, spectrum, time series, etc.) are to be defined. Thus DAL is oriented on support of limited number of 'built-in' data types by respective wrappers (adapters) implemented as services. No visible attempts of IVOA to provide general architectural DAL facilities for querying services established above multiple similar sources (e.g., image archives) and to compose such queries with data queries (e.g., for heterogeneous catalogs) are known.

In contrast, the mediator layer is more concerned with user-defined, arbitrary subject mediator types that can be queried (queries include a composition of data and functional terms). It is important that wrappers constitute important part of the mediator architecture located at the bottom level of the mediators and also designed as Web (Grid) services.

Due to the above, DAL might be considered as an important step to the mediator architecture proposed for RVO. At the same time, an extension of DAL can be considered to encompass the subject mediator architecture.

SkyQuery vs Mediator Querying

SkyQuery [BSKYQ] is designed according to the so called Global as View approach. Global schema is provided as a view above heterogeneous data sources. To make the design even simpler, the global schema is formed as a multibase schema in which each table that corresponds to a table in some catalog has a respective uniform schema representation. So the user sees the data base as multiplicity of tables. Each time when new catalog is to be registered in SkyQuery, the global schema is to be extended. Another unpleasant consequence of the approach is that it is not scalable. Even

integration of 100 catalogs will create serious problem for the user: the global schema will be hardly operable.

In contrast, the Subject Mediator exploits Local as View approach. Sources are considered as materialized views over the mediator schema. Mediator schema reflects definitions related to the subject domain and is not changed with addition/removal of sources registered at the mediator. The mediator schema reflects concepts, data structures and behaviors corresponding to understanding of the respective subject domain by the user. Conceptually no scalability problem arises for the user when new sources are registered at the mediator.

Mediator querying and sources registering might be an issue for the IVOA architecture extension.

IVOA metamodeling vs Mediator metamodel

IVOA recommendations are inconsistent w.r.t. using of metamodels, e.g. UML and object-oriented specifications are used by the Data Modeling and DAL groups, alongside with a relational model used by VO Query Language Group. Probably, one of the reasons for such inconsistency is an absence of the architectural requirements that would strictly interconnect data models, DAL protocols, database schema and query languages into a consistent, uniform architecture. Astronomical artifacts by their nature are object-oriented. Various services related to them can also be well defined in such paradigm. Specifically this is important to understand for designing a researcher environment that should be natural to formulate the researcher problems. Basic ingredients of the problem formulation (besides processes) include data filtering and data analysis. Both should be expressible in natural way. Using obsolete SQL standard for data filtering (ADQL) looks as a short-sighted decision. A necessity to use alternative facilities by Data Modeling and DAL groups confirm this.

The mediator architecture, on the contrary, enforces to consider data models, DAL protocols, database schema and query languages as well as problem formulation models as highly interconnected and well agreed facilities. This is why selection of the mediator canonical model is an issue of prime importance: the model should be natural for all the ingredients mentioned above. It might be not easy to define subject mediation architecture consistently applying IVOA recommendations that look sometimes as not consistent enough.

10.2 An example of a subject mediator for a specific problem class

During several years the Big Trio project [PBTRI] is carried out in SAO RAS headed by the academician Parijskij Yu. N. The main project task is a

distant radio galaxy search in the sky strip investigated in the “Cold” deep survey with the RATAN-600 in 1980 and getting maximal information about the objects. The project is called “Big Trio” because of the three large instruments used for deriving of observation data. RATAN-600 is a source of primary information about radio objects, VLA (NRAO, USA) is used for getting radio images and perfect radio source coordinates and 6-m telescope is utilized for optical identification and spectroscopy. Distant galaxy candidate from radio source lists and catalogs is derived with tested selection methods by certain radio source parameters. The candidates for distant galaxies were selected from RC catalog objects applying these methods [KDORC]. Similarly to the other research groups involved into the same problems, the following parameters of radio sources were used:

1. A slope of radio source spectrum. Steep spectrum sources were selected (spectral index is in range 0.9 – 1.2);
2. Flux density level. RC catalog objects on the average have fluxes about 100 mJy. Flux densities mean level of the RC catalog fall within inflection area of normalized curve logN-LogS (source number – flux density). Apparently this area may include large number of distant objects;
3. Morphological type. The objects of FRII morphological type are selected (these are powerful radio galaxies which could be seen on the large distance);
4. Angular size. Large angular size radio galaxies were not registered for considerable redshifts. Their usual angular sizes are in range 1 arcsec to 1 arcmin;
5. Proximity of radio and optical luminosities. The property is used for estimation of reliability of the 6-m telescope optical identification of radio sources and in addition for separation of radio galaxies and quasars.

Initially all RC catalog sources with spectral indexes near 1 and more were picked out (about 100 objects). Then radio maps, accurate coordinates for optical identification, morphology and angular sizes were obtained by VLA for the sample objects. Special observations were performed for the sources or their images were fetched from the FIRST and NVSS radio surveys. All objects were identified by the 6-m telescope CCD-images.

The necessary and often sufficient condition for the optical object correlation with radio source is positional coincidence by coordinates but there are other factors which may increase or decrease probability of identification. If coordinate precision of radio and optical catalogs is high then positional coincidence gives high degree of identification reliability.

Actually, the reliability of radio/optical identification is a difficult problem depending on an astrometric precision of two data sets and structure of optical candidates and radio source morphology.

Photometric redshift estimations and ages of galaxy stellar population were carried out with Spectral Energy Distribution (SED) models and the 6-m telescope B, V, R, I observations. The method precision by our estimation is not worse than 30%. Old stellar population availability was confirmed among objects with high redshift.

10.2.1 The mediator schema

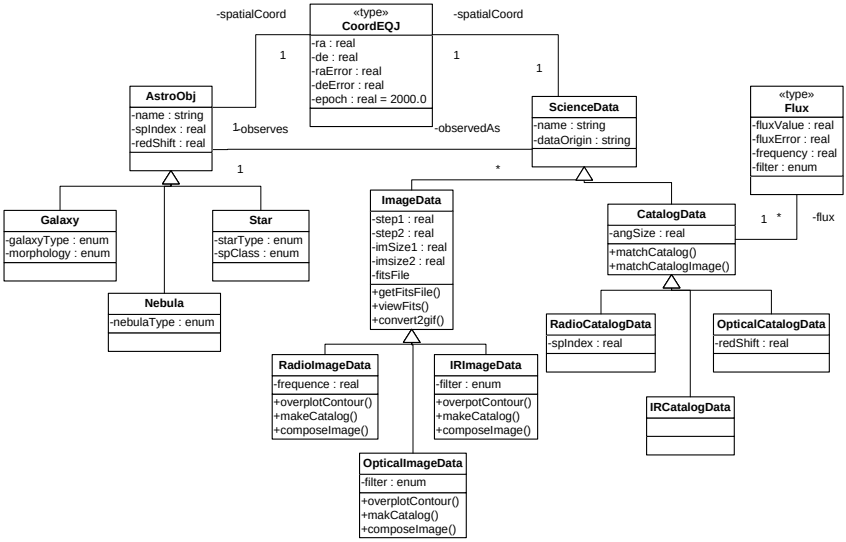

Figure 6. The mediator schema

Fig. 6 presents UML specifications of types of the mediator schema for distant galaxy discovery. The astronomical objects (galaxies, stars, nebulas) and observed scientific data (images, spectra, catalogs) are distinguished. The relationship between astronomical objects and observed data are established during processing and analysis of observed data for a given object in different spectral ranges.

Each schema element (class, types, attribute, function) is annotated with the relevant UCD descriptor. Such annotation for some attributes of the mediator looks as follows:

```
RadioCatalogData.name    [UCD: ID_MAIN]
RadioCatalogData.spIndex [UCD: SPECT_SP-INDEX]

CoordEQJ.ra          [UCD: POS_EQ_RA_MAIN]
```

```
CoordEQJ.raError     [UCD: ERROR]
CoordEQJ.de          [UCD: POS_EQ_DEC_MAIN]
CoordEQJ.deError     [UCD: ERROR]
CoordEQJ.epoch       [UCD: TIME_EQUINOX]

Flux.fluxValue       [UCD: PHOT_FLUX_RADIO_3.9G]
Flux.fluxError       [UCD: ERROR]
```

10.2.2 The process of distant galaxy candidates discovery

The process of investigation of radio galaxies with steep spectra (based on the methods proposed by Parijskij Yu. N. for the Big Trio project) consists of the following steps:

1. select radio sources which spectral index is confined in a given range [i1,i2];
   - 1.1 select radio sources which contains flux data for at least 2 frequencies;
   - 1.2 calculate spectral index i for selected radio sources;
   - 1.3 give a list of radio sources which spectral index is confined in the range [i1,i2];
2. select radio sources which angular size is less then a given value d1 and flux value for a given frequency f1 is in the range [s1,s2];
3. find optical sources matching by coordinates (admissible w.r.t. coordinate and size of sources tolerances) with selected radio sources;
4. get a list of radio sources for which the matching optical sources were found;
5. on a user request, for a given pair – radio source and matching optical source;
   - 5.1 show optical image of the source with over plotted intensity contours of radio image;
   - 5.2 show radio image of the source with over plotted intensity contours of optical image;
6. request user confirmation of matching of the selected radio and optical sources;
7. calculate color values c1 and c2 with a given filters for selected optical sources;
8. select sources which calculated color values c1 и c2 are in the ranges [a1,b1] and [a2,b2] respectively;
9. get a list of sources– candidates for distant galaxies.

10.2.3 Resources to be registered in the mediator

Initial set of catalogs and surveys to be registered at the mediator includes:

- RC catalog – catalog of radio sources;
- FIRST – survey and catalog of radio sources;
- NVSS – survey and catalog of radio sources;
- 2MASS – two micron all sky survey;
- SDSS –survey of one-quarter of the entire sky, in 5 spectral ranges from ultraviolet to infrared;
- SAO RAS archive, direct images in U, B, R filters.

10.2.4 Resources registration at the mediator

During the registration a resource class is modeled as a set of instances (objects) of a mediator class instance type and the description of the source in terms of the mediator schema specifies the constraints on the class instances to be admissible for the subject mediator. The registration process includes: ontologically-based (UCD-based) reconciliation of the application contexts of the registered resource and that of the mediator; expression of each resource class in terms of the mediator classes.

RC catalog

Fig. 7 shows a fragment of RC catalog specification. It includes specification of *RCCatalog* type and of its attributes.

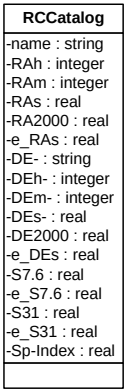

Figure 7. A fragment of UML specification of RC catalog

Each attribute may be annotated with the respective UCD:

```
name        string          [UCD: ID_MAIN]
RAh         integer
```

```
RAm         integer
RAs         real
RA2000      real          [UCD: POS_EQ_RA_MAIN]
e_RAs       real          [UCD: ERROR]
DE-         string
DEh         integer
DEm         integer
DEs         real
DE2000      real          [UCD: POS_EQ_DEC_MAIN]
e_DEs       real          [UCD: ERROR]
S7.6        real          [UCD: PHOT_FLUX_RADIO_3.9G]
e_S7.6      real          [UCD: ERROR]
S31         real          [UCD: PHOT_FLUX_RADIO_1G]
e_S31       real          [UCD: ERROR]
Sp-Index    real          [UCD: SPECT_SP-INDEX]
```

Establishing relevance between mediator schema and RC Catalog schema elements

Using UCDs, we can establish relevance between elements in the mediator and RC catalog schemas. If for a mediator attribute there is no directly relevant RC catalog attribute, then we construct, if possible, a concretizing (or conflict resolving) function. This function defines how a mediator attribute can be expressed as a function of the resource attributes.

The set of relevant elements of mediator schema and RC Catalog schema looks as follows:

```
RadioCatalogData ~ RCCatalog
name ~ name
spatialCoord.ra ~ calcRA2Deg(Rah, Ram, RAs)
spatialCoord.raError ~ calcArcsec2Deg(e_RAs)

spatialCoord.de ~ calcDE2Deg(DE-,DEh,DEm,DEs)
spatialCoord.deError ~ calcArcsec2Deg(e_DEs)

flux.fluxValue ~ mJy2Jy(7.6)
flux.fluxError ~ convToFluxJyErr(7.6)
flux.fluxFrequency ~ sm2MHz(7.6)
flux.fluxValue ~ convToFluxValueJy(31)
flux.fluxError ~ convToFluxErr(31)
flux.fluxFrequency ~ convWavelenghToFrequenceMHz(31)

spIndex ~ Sp-Index
```

The conflict resolving functions look as follows:

```
calcRA2Deg (h,m,s) = (h+m/60+s/3600)*15
calcArcsec2Deg(s) = s/3600
calcSec2Deg(s) = s/3600*15
calcDE2Deg(sign,d,m,s) = {deg  = d+m/60+s/3600; if sign = "-
" then deg = -deg}
mJy2Jy (i) = i*0.01
```

10.2.5 Example of the mediator queries for different steps of the process of distant galaxy candidates discovery

Object query language has been used for these examples.

Select radio sources containing flux data for at least 2 frequencies:

```
select r.spatialCoord.ra, r.spatialCoord.de,
r.spatialCoord.raError, r.spatialCoord.deError, r.flux,
r.angsize from RadioScienceData r where r.flux.size() > 1
```

Calculate spectral index for selected radio sources:

```
select temp1, spind: t.calcIndex() from temp1 t
```

Get a list of radio sources with spectral index in a given range [ind1, ind2]:

```
select temp2 from temp2 t where t.spind between ind1 and
ind2
```

Select radio sources with angular size less than a given value d and flux value for a given frequency freq1 in the range [i1,i2]:

```
select temp3 from temp3 t where t.angSize < d and
t.flux.frequency = freq1 and t.flux.fluxValue between i1 and
i2
```

Find optical sources matching by coordinates (admissible w.r.t. coordinate and size of sources tolerances) with selected radio sources;

```
select oCoord: o.spatialCoord, oFlux: o.flux from temp4 t,
OpticalScienceData o where t.match(o)
```

Get a list of radio sources for which the matching optical sources were found:

```
select o.Image, r.Image from temp5 t, RadioImageData r,
OpticalImageData o where t.overplotContour (r.Image,o.Image)
```

Calculate color values with a given filters f1 and f2 for selected optical sources:

```
select oFlux, oCoord, colorind: t.calcColorInd(f1,f2) from
temp6 t where t.flux.filter = f1
```

Select sources for which calculated color value is in the range [a1,b1]:

```
select oCoord from temp7 t where t.colorind between a1 and
b1
```

10.2.6 Example of the process (workflow) specification for problem solving by means of a mediator

An example of a workflow for distant galaxy candidate choice from radio source lists is shown on Fig. 8.

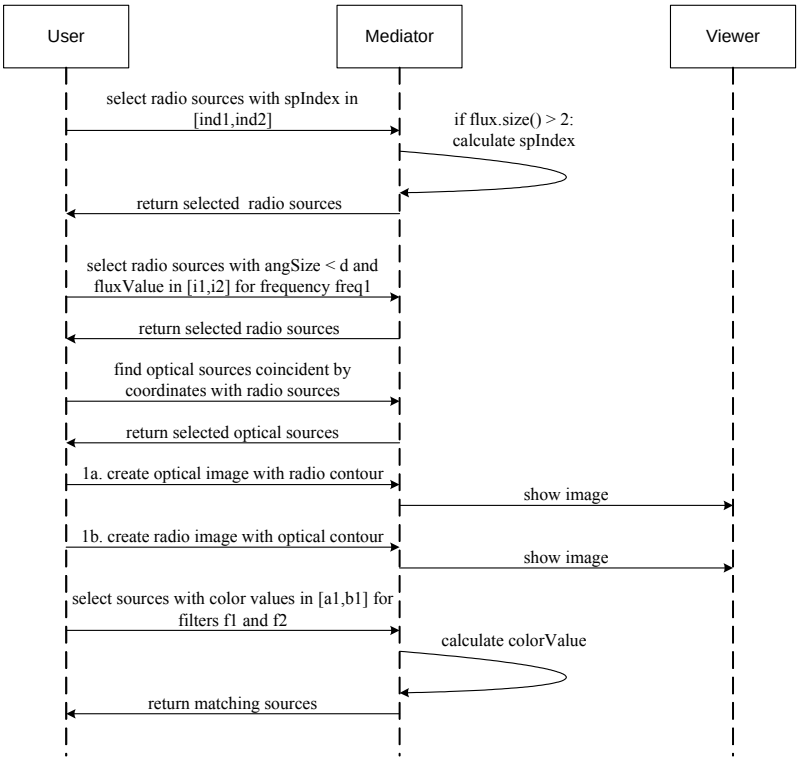

Figure 8. An example of workflow specification

11 Information infrastructure of the RVO

11.1 Matching of the IVOA standards to the RVO activities

To analyze matching and sufficiency of the IVOA standards to the RVO, the analysis of the activities that should be supported by the RVO has been undertaken. The result of the analysis is presented as a list of the RVO activities that follows. It is assumed that all activities defined in this section will be supported by web services and will be accessible from programs.

Data centers (resources) development

Data centers activities consist of publishing data, metadata and services to the VO and providing a research infrastructure through the adoption and application of processing and storage facilities supplied by the VO technical infrastructure (e.g., Grid). Data centers should provide astronomers with easy, long term access to observation (simulation) archives and expert services to the community. These activities are decomposed further in more details.

- Archive access provision

IVOA DAL standards for uniform access to archives data that may have heterogeneous representations by different data providers is related to these activities. Currently DAL WG provides standards for development of the required wrappers and web services for Simple Image Access (SIA) and Simple Spectral Access (SSA). These core services provide simple access patterns, e.g. an image mosaic, assembled from parts, centered on a given point, in standard orientation and at a specified scale. Another simple access activity is known as a Cone search that is also supported by the core service.

Important archive access activity consists in building a catalog of objects from astronomical images.

Another simple activities related to the archive access include various functions for format transformation and viewing (such as creating and editing FITS files, converting FITS images to other image format, converting FITS tables to VOTable format, merging together FITS images using any arbitrary astrometric projection, etc.).

- Catalogs creation and support

Catalog creation activity includes design of the catalog data model as of a database, its physical design and its population with data. Due to the fact that IVOA does not plan any standard for the catalog modeling, defining the semantics of the catalog data model (partially the semantics can be defined using UCDs and/or ontologies) is very important part of the catalog design.

- Resource metadata registry support

Resource metadata are assumed to be defined by the resource provides (e.g., observatories, simulations). Minimally a local publishing registry should be supported so that its content could be harvested by a searchable VO registry.

- Data center metadata registry support

It is assumed that the data center metadata registry should be local searchable registry that attempts to contain records of all resources known to the center. Such registry gathers its descriptions from across the network through harvesting.

- Data center catalog warehousing

Important function of data centers consists in collecting published astronomical catalogs from the region and providing access to their tables. Query tools allow the user to select relevant data tables and to extract and format records matching given criteria. Development of such warehouses and their support is an important activity of the data centers.

Robotic telescopes

- Robotic telescope support

It is assumed that a robotic telescope provides facilities for submitting observation requests and requested analysis of the reduced data returned from the observation. Requesting additional data and follow-up observations from the telescopes on the network are also possible. It is assumed that in the RVO infrastructure the robotic telescopes will be represented by the respective web services.

Inclusion of data resources (data centers) in VO

- Metadata integration

This activity consists in creation and support of the full metadata registry that attempts to contain records of all resources known to the VO. Full metadata registry is formed by harvesting of searchable registries around the world (region – e.g., Europe).

- Providing unified semantics

To be included, each resource should be provided with semantic description of its constituents (minimally names used are to be annotated with UCDs).

Development of an ontology for the astronomy (or its branches) is a far reaching goal. The IVOA document “A unified domain model for astronomy” might be considered as a step to the reference ontology in the community.

- Creation of the VO nodes for the integrated search

Each resource is to be supplied with the required interface to participate in VO as a queryable VONode. Currently basic SkyNode is the minimum IVOA

SkyNode Interface that allows one to send queries to a resource using ADQL. This is just one step up from the cone search. Full SkyNode capabilities allows the node to be used as an intermediate node in calculation of a query: it should be able to request and manage a subplan of query execution and calculate cross-identification for different catalogs involved in a query.

It is assumed that Data Centers should be provided with Full SkyNode capabilities.

Discovery of resources (services)

- Metadata-based discovery in registries

Automated resource discovery is an important data center activity (e.g., such requests as find the archives, which have optical and infrared data about a certain part of the sky, or find archives which have spectra at higher than 1 Angstrom resolution are assumed to be typical for such activity).

- Metadata-based discovery in warehouses

The discovery activity in the catalog warehouses should lead to finding of catalog tables relevant to the search.

- Providing access to metainformation of catalogs and services

Access to metainformation is important activity required for selection of resources relevant to a specific problem. Basic metadata are defined by the resource metadata for registries defined by IVOA. Additional metainformation can be obtained using the VONode interface. The metainformation that is defined currently by IVOA is not complete for justifiable selection of relevant resources.

Support of collaboratory dataspaces (MySpace)

- Management of dataspaces

An activity of organizing personal and group data spaces during problem solving. MySpace provides astronomers with data storage that is not necessarily at their home institution and which they can use to store the result sets from queries, workflows, private datasets, etc. related to their work. Users will be able to interact with their holdings in MySpace as easily as if they were directly on a local machine. A user may publish any of his/her MySpace tables to the groups of which he is a member. The MySpace concept provides an area where 'in progress' data and results can be accessed securely and remotely by user defined working collaborations. The data will be stored in standard VO format with standard metadata. Resource could be used in other jobs as simply as data sets from data centre (easier in fact as the data will already be in standard format).

Integrated access to catalogs

- Single catalog search

This activity is supported by the Cone search and ADQL-based search in a single catalog.

- Integrated search in catalogs

Search in the integrated catalog of VO should be provided. ADQL and cross-identification facilities are considered as minimal requirements. Modeling of the integrated catalog for VO is an open issue. Current SkyQuery technology looks as not scalable. One of the possibilities for RVO is to apply mediation technique and apply integrated search in the problem domain of a mediator. OGSA DAI DQP might be considered as another technology (though also not scalable)

Data model development for subject domains

This activity is of top importance not only for the IVOA, but also for the community. It is doubtful that the conceptual data model for the astronomy as a whole could succeed. At the same time, identification of subject domains in astronomy relevant to active areas of research and development of conceptual models for the domains might be useful activity that might complement the current activity of the IVOA DM Group.

Development of mediators for problem classes

- Mediators definition

The activity consists in the identification of a problem class worth of design a mediator and definition of the mediator applying the canonical model.

- Registration of relevant resources and services at the mediator

The activity consists in discovery and registration in mediator of relevant information sources (data sources and/or services). Discovery of relevant resources and services is based on matching of their definitions of metadata, ontologies, structure and behavior with the respective definitions in the mediator. Resources are registered as materialized views over virtual classes of the mediator. During the registration, the required mediator functions are attempted to be implemented by composition and reuse of existing services.

- Problem solving applying the mediators

For this activity the mediators should provide convenient user and program interfaces.

VO data processing and analysis

- Type specific data analysis

This research-oriented activity consists in applying type specific data analysis (e.g., of images, spectra, time series, etc.). Specific tools are required for that. Composition of facilities for data search (discovery) and analysis should be provided.

- Object kind specific data analysis

This research-oriented activity consists in applying object specific data analysis (e.g., galaxy oriented). Specific tools are required for that. Composition of facilities for data search (discovery) and analysis should be provided.

- General data analysis

This research-oriented activity consists in applying general kinds of analysis (e.g., data mining (predictive and descriptive features), statistic analysis)

Development of theoretical (simulation) models

- Simulation activities

Activity for developing simulations which encode current understanding of the world model, of the clustering evolution of dark matter, of the physics of galaxy formation, etc. Tools should be developed for that.

- Publishing simulation results

An activity for simulation results publishing should be supported. Sets of simulated objects should be published as conventional resources in the metadata repositories to make them available to various groups of researchers, educators and students around the world.

- Analysis of simulation results

An activity of simulation results analysis and comparing them with observational data should be supported by specific tools. Many of the outstanding problems in cosmology are inherently statistical, either studying the distributions of typical objects or finding the atypical objects. Exponentially growing astronomy data volumes impose serious new requirements for the algorithms. Computational Grid should provide an infrastructure for this class of problems.

User access to VO

- Portals

An activity of various user communication with VO during their research/educational facilities should be supported. Respective kinds of portals should be provided for different categories of VO users/tasks. Classification of VO users/tasks is required for that.

Development of digital libraries for the educational resources in astronomy

- DLAE metadata registries

DLAE metadata registries definition and support is a specific activity. Metadata models for DLAE are to be developed and respective metadata registries should be created. DLAE and VO resources should be interconnected.

- Resource selection/development

Web contains many resources in astronomy worth of their inclusion into the DLAE. Specific activity for the resources selection/creation is required to populate the DLAE.

- DLAE usage

Activities based on DLAE extend education and research in astronomy. These activities should be supported by the respective user interfaces.

Supervision

- Data access management (authorization)

Data access management activity (such as authorization) should be supported.

- Workflow management

Important part of research-oriented activity applying VO consists in defining processes of steps required to reach the objectives of research.

- Job control

Important part of computer application activity consists in supervision of jobs consisting of sequences of services (programs) to be executed.

Provision of the general infrastructure

Important activity of the VO design and development consists in provision of general infrastructure for the VO support (such as Web services or Grid).

Correspondence of the IVOA standards to the RVO activities is established by the following table.

Activities	IVOA standards	Standard required
Data centers (resources) development		
Archive access provision	DAL SIA and SSA VOTable	
Catalogs creation and support	Basic SkyNode	

Resource metadata registry support	IVOA Metadata IVOA registry interface	
Data center metadata registry support	IVOA Metadata IVOA registry interface	
Data center catalog warehousing	DAL Cone Search ADQL	
Robotic telescopes		
Robotic telescope support		Robotic telescope interfaces
Inclusion of data resources (data centers) in VO		
Metadata integration	IVOA Metadata IVOA registry interface	
Providing unified semantics	UCD	
Creation of the VO nodes for the integrated search	Basic SkyNode Full SkyNode	
Discovery of resources (services)		
Metadata-based discovery in registries	IVOA Registry Interface ADQL	
Metadata-based discovery in warehouses		Catalog warehousing for data centers
Providing access to metainformation of catalogs and services	Basic SkyNode	
Support of collaboratory dataspaces (MySpace)		
Management of dataspaces		Dataspace standards

Integrated access to catalogs		
Single catalog search	DAL Cone Search ADQL, VOQL VOTable	
Integrated catalog search	SkyQuery ADQL, VOQL VOTable	
Development of mediators for problem classes		
Mediators definition		Canonical information model
Registration of relevant resources and services at the mediator	IVOA metadata UCD	Resource Node Interface (genera-lization of SkyNode)
Problem solving applying the mediators		
VO data processing and analysis		
Type specific data analysis		
Object kind specific data analysis		
General facilities		
Development of theoretical (simulation) models		
Simulation activities		
Publishing simulation results	Simulation DM	
Analysis of simulation results		
User access to VO		
Portals		

Development of digital libraries for the educational resources in astronomy		
DLAE metadata registries		DLAE metadata standard(s)
Resource selection/development		
DLAE usage		
Supervision		
Data access management (authorization)		
Workflow management		Workflow process model standard
Job control		
Provision of general infrastructure		
Web services	W3C standard	
Grid	OGSA DAI, WSRF	

The following directions are identified for the first order additional standards development:

- Catalog warehousing for data centers;
- Dataspace standards;
- Canonical information model for the mediation facilities intended for specification of subject domains (including ontologies) and problem formulation;
- Resource Node Interface for registration in mediators (possibly as a generalization of SkyNode Interface);
- DLAE metadata standard(s);
- Workflow process model standard;
- Robotic telescope interfaces.

11.2 RVO infrastructure overview

11.2.1 Basic principles for the RVO infrastructure

New discoveries require federating as many observations as possible. The observational data sets are in distant places around the world. Therefore VO is a decentralized system.

Basic RVO infrastructural principle is to represent the architecture as a network of interoperating web services (Grid services as soon as suitable OGSA DAI or WSRF standard will mature). A well-defined set of core services implemented by most of the data providers will ensure a similarity in the basic functions and enable their automated integration at higher levels. Then layers of much more complex processing steps can be imposed above the core.

Thus a multilevel hierarchy of services is the basis for the RVO architecture. The handling of remote and virtual data sources should be provided. The core will be set of simple, low level services that are easy to implement even by small projects. Thus the threshold to join the VO will be low. Large data providers may be able to implement more complex, high-speed services as well. The services can be combined into more complex compositions that talk to several services, and create more complex results.

Move processing to the data is another principle motivated by large volume of the data and data intensive character of VO applications. After the first few steps of processing the output volume becomes significantly smaller (e.g. extracting object catalogs). In many cases the data will not only be remote, but it does not even exist at the time of the request: it may be extracted from a database with a query created at that moment.

Building software for more and more projects is a main challenge for the VO development and evolution. Modular architecture that encourages code reuse and composition is another guiding principles for the RVO infrastructure.

Due to existence of thousands of astronomical data resources (e.g., catalogs), conventional practice of applying global as view approach to data integration in the VO projects (e.g., SkyQuery) looks as not scalable. Another approach assumes creation of mediators supporting interaction between a researcher and relevant data sources and services through a subject domain description for a class of problems. Emphasizing subject mediators to support representation and access to various subject domains in astronomy is a basic RVO principle.

11.2.2 The RVO layered infrastructure

Conceptually the information infrastructure of the RVO is considered to be multilayered (Fig. 9). Each layer of the infrastructure includes information entities, access services, data analysis and user support facilities intended for

the information entities of the respective layer. We assume that the infrastructure should support the RVO activities considered in the previous section. Primary information sources (archives, simulation results, robotic instruments and various publications related to astronomy) form the ground layer of the infrastructure.

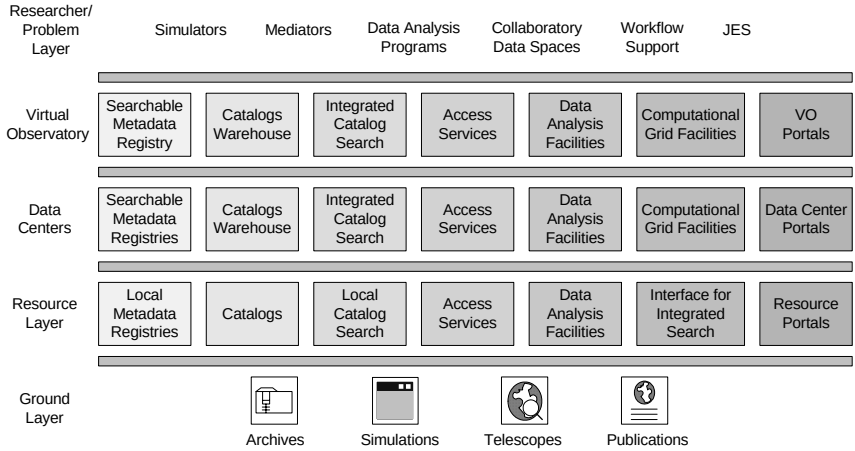

Figure 9. The RVO layered infrastructure

The closest layer to the Ground, the Resource Layer (RL) gives to primary information source providers an ability to publish their holdings and make their services available for the users and higher layers of the RVO infrastructure. RL contains components providing for:

- Catalogs creation, storage and access;
- Single (non-integrated) catalog search;
- Publishing metadata registry facilities providing for metadata-based resource discovery in registries;
- Access to the Ground layer sources by means of the respective services (by unifying wrappers over images and spectra constructed in accordance with the IVOA DAL standards or specific services for simulation or data analysis). This is shown on Fig. 10;
- Data processing and analysis facilities including low level functions (such as type specific data analysis, astronomical object kind specific data analysis, as well as various functions of data transformation between various primary data sources formats);
- Creation of the VO nodes for the integrated search on the higher layers of the infrastructure;
- User access to the Ground and Resource layer data and services (portals);

- Application program interfaces to various functions of the Resource layer.

The components of RL are properly interconnected horizontally, they get access to the relevant resources of the ground layer and provide the required interfaces for the upper layers. For catalogs of RL the full IVOA SkyNode interfaces are to be provided.

Next above the Resource layer is the Data Center Layer (DCL). DCL introduces additional level in the RVO information organization hierarchy. Data Centers create National Nodes as the integrating facilities for the National and the European levels. Data Centers may be created based on a regional principle, orientation on specific kinds of astronomical objects, or other orientation. They should conform to common protocols and standards. Graphically the DCL components are similar to that of the RL layer (Fig. 9).

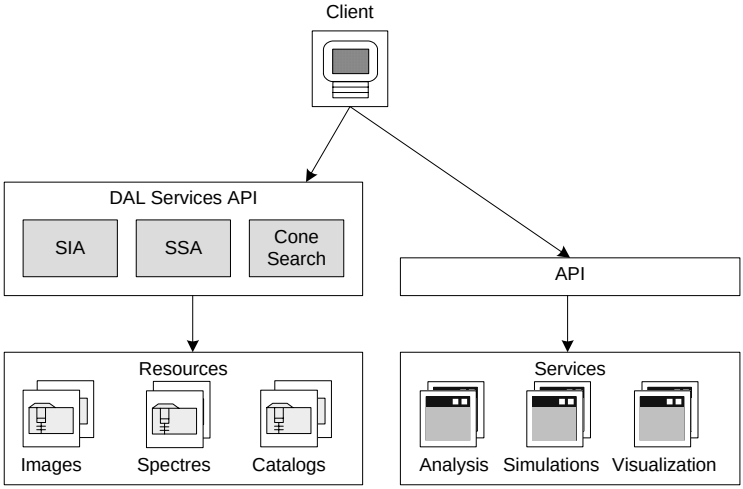

Figure 10. Access services provision

The following differences between these layers worth of noting:

- DCL metadata registry facilities possess the searchable capabilities in accordance with the IVOA standard. Their formation is based on applying of the metadata harvesting with the OAI-PMH facilities (Fig. 11);
- DCL supports facilities for astronomical catalogs published at the RL warehousing, providing a complete library of catalogs and data tables in an area of the respective Data Center;
- DCL provides the integrated catalog search applying technique similar to that of SkyQuery (it is assumed that the catalogs involved into the integrated search provide the full IVOA SkyNode interfaces). It is

assumed that the DCL catalog search facilities possess the required capabilities to be involved into the integrated search facilities of the VO Layer;

- Data analysis services of the DCL Layer provides higher facilities for research comparing to the RL layer;
- DCL provides Grid-enabled computational facilities for computationally intensive research (like simulation or statistical analysis).

The VO Layer (VOL) architecturally is similar to DCL and provides final layer of the RVO information integration hierarchy. The intention of this layer is to provide facilities to access informational and computational resources available in frame of the International VO.

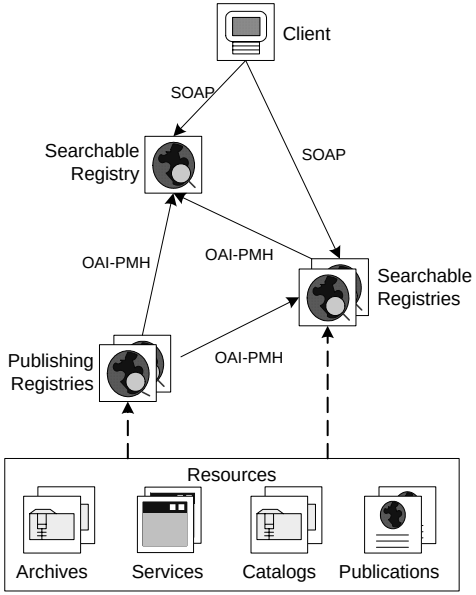

Figure 11. Searchable metadata registries at the Data Center and Virtual Observatory layers

The Research/Problem Layer (RPL) is intended to support problem solving by the researchers using VO facilities at all layers. Data analysis programs, simulators and mediators are different kinds of facilities that can be developed in course of research. Architecture of subject mediators is discussed in the next section. Important ingredients of RPL include facilities for workflow definition and management as well as for job execution support. Collaboratory dataspace management is provided for user's (and user group) own data organization within the RVO.

Nonrestrictive technological hints for implementation of a layer of the RVO infrastructure is shown on Fig. 12. Further concretization of technological facilities for a layer is planned during planned technical development of the RVO project.

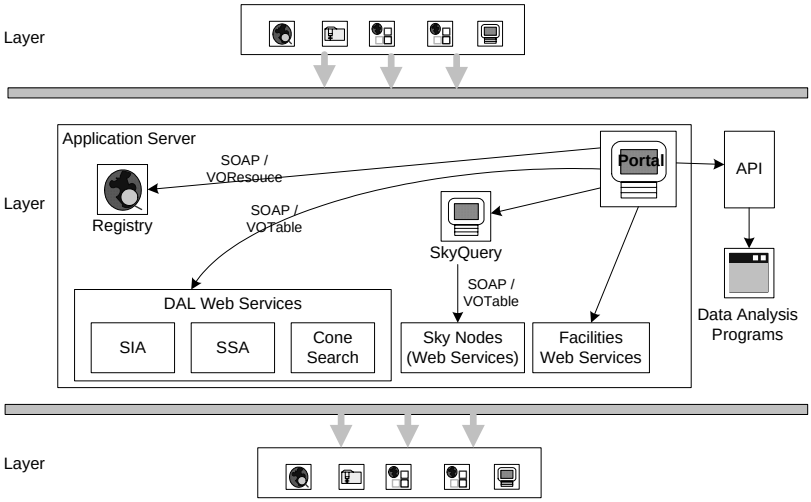

Figure 12. Technological facilities for implementation of a layer of the RVO infrastructure

11.2.3 Components of subject mediators

Due to the fact that subject mediators constitute novel component of RVO comparing to other known VO infrastructures, their architecture is explained in more details.

To provide for interoperability of heterogeneous information collections, it is required to establish a global, uniform view of the underlying digital sources and services. It is assumed that specific, intermediary layer is formed by *mediators* providing a uniform schemas and query interfaces to the multiple data sources and services to free the users from having to locate the relevant sources, query each one in isolation, and combine manually the information from the different sources. Thus, the mediators are intended to support an interaction between a researcher and relevant data sources and services through a subject domain description for a class of problems. A subject domain definition for a class of problems roughly is subdivided into the three interrelated parts: concept space; results of experiments, observations, measurements, simulations, intermediary results; computational models. *Subject mediators* are emphasized that support representation and access to

various subject domains in astronomy. The mediator's metainformation is intended to be shared by information consumers, source providers and subject mediators.

Mediators include the following components

1) A *canonical information model* is used for the mediator definitions making possible to query subject mediator content and compute the result. The mediator's canonical information model should be powerful enough for uniform equivalent and complete representation of various data sources and services. For the canonical model of a mediator a comprehensive object model is proposed. The canonical model [KCSYN] provides support of wide range of data – from untyped data on one end of the range to strictly typed data on another. Typed data should conform to the abstract data type (ADT) specifications prescribing behavior of their instances by means of the type operations. ADT describes interface of a type whose signature defines names and types of its operations. Classes are collections of instances of respective types. Subtyping relationship is supported by the model.

Formulae in the language are used to specify queries, rules and constraints. To specify formulae a variant of a typed (multisorted) first order predicate logic language is used. Predicates in formulae correspond to collections (such as sets and bags of non-object instances), classes treated as set subtypes with object-valued instances and functions. ADTs of instance types of collections and classes should be defined. Rules look as:

$$q(v/T_v) :- C_1(v_1/T_{v1}), \ldots, C_n(v_n/T_{vn}), F_1(X_1,Y_1), \ldots, F_m(X_m,Y_m), B$$

where in the body (having SPJ semantics) C_i is a class predicate, F_i is a functional predicate and B is a condition. Rules can be combined in program blocks applying conditionals and iterators if required. Note that rule bodies can be easily converted into SQL-like notation.

To define process types (to support workflows), the canonical model is augmented with scripting facilities based on the colored Petri nets.

2) The mediator metainformation registry system that uses the canonical information model constructs to hold the mediator definition and link with it diverse contexts and representations of metainformation related to heterogeneous sources and services registered at the mediator.

3) Tool for discovery and registration of relevant information sources and services at the mediator.

Each mediator supports the process of systematic discovery and registration of sources uniformly expressing their definitions in terms of the mediator. The process of registration is assumed to be semi-automatic.

Discovery of relevant data types (classes) in sources is based on three models: metadata model describing the information sources (according to the IVOA standard), ontological model defining concepts of a subject domain

(generally, ontological model is a subset of the canonical information model) and canonical model providing for definition of structure and behavior of objects of the subject domain and of the source.

The tool for discovery and registration supports: 1) metadata search, 2) for the sources found the reconciliation of their ontological context with that of the mediator, 3) identification of the mediator classes ontologically relevant to the registered source class, 4) mediator type refinement by composition of relevant source types, 5) designing specification of source classes as materialized views over the virtual mediator classes applying constraints given in the terms of the mediator classes.

The main distinguishing feature of the compositional development method applied for the registration is a creation of compositions of component specification fragments refining specifications of requirements. Refining specifications obtained during the compositional development can be used anywhere instead of the refined specifications of requirements without noticing such substitutions by the users. Note that in this process, the mediator classes instance types are refined by composition of the source classes instance types. This is a natural direction supporting the process of rewriting a mediator query in terms of specific sources. This approach is valid also for reuse of functions (and their compositions) provided by services known to the VO.

The registration approach is intended to cope with a dynamic, possibly incomplete set of sources. Sources may change their exported schemas, become unavailable from time to time. To disseminate the information sources, they should be registered at respective subject mediators. Such registrations can be done concurrently and at any time.

To map heterogeneous source models into the canonical one, the commutative mapping method is applied. To preserve information and operations of types of a specific data model while mapping them into the canonical types the commutativity of two mapping diagrams (data type state and data type behavior diagrams) should be established. The required state-based and behavioral properties of the mappings lead to a proof that a source data model is a refinement of its mapping into the canonical data model.

4) Rule rewriter

The rewriter transforms rules used in queries and programs expressed in the canonical model over the mediator schema into the rules over the sources registered at the mediator.

5) Planner

Planner imposes an order of concurrent execution of the program at different source nodes where the respective sources and/or services reside and exchange of intermediate results of the subquery execution.

6) DBMS of the mediator

The DBMS is used for execution of the residual subquery that provides final composition of results obtained from various sources.

7) Executive

Coordinates the mediator components collaborative work during the mediator programs execution.

8) Interface

Provides an interface for communication with clients.

The mediator architecture is represented on the Fig. 13.

It is important to note that sources and services shown on the figure and registered at the mediator can belong to any layer of the RVO infrastructure.

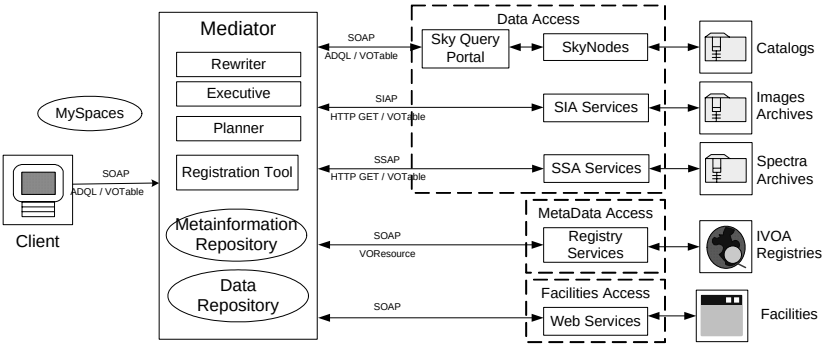

Figure 13. Architecture of a mediator

11.3 Existing components (prototypes) that can be used in the RVO infrastructure

In this subsection an analysis of existing software components that can be suggested for implementation of the RVO information infrastructure is given.

11.3.1 Data centers (resources) development

Archive access provision

IVOA FITS package

The IVOA FITS package (http://skyservice.pha.jhu.edu/develop/vo/ivoafits/) is a set of classes which can be used by any Java application to load and view FITS images. It is currently used to provide image viewing functionality in VO Enabled Mirage. FITS file data loading uses Tom McGlynn's nom.tam.fits package.

conVOT

conVOT (http://vo.iucaa.ernet.in/~voi/conVOT.htm) is a tool for converting ASCII or FITS tables to VOTable format. For ASCII files, it supports both ASCII files with column delimiters and ASCII files with fixed width columns. For FITS files, it supports FITS ASCII and Binary tables. Accessed as Java standalone application. conVOT uses nom.tam.fits library for reading FITS tables. nom.tam.fits is developed at Heasarc. conVOT uses VOTable Java Streaming Writer for writing the data in VOTable format. The VOTable Java Streaming Writer is developed as part of Virtual Observatory India initiative.

VOTable Java Streaming Writer

A number of tools are available to deal with data in the VOTable format. These very useful packages cannot be used with data in other formats (like ASCII or FITS) and therefore there is the need to convert such data to the VOTable format. There is also the need to convert data streams produced in non-VOTable format by various applications, to the VOTable format.

The "VOTable Java Streaming Writer" (http://vo.iucaa.ernet.in/~voi/votableStreamWriter.htm) acts on a data array in memory to convert it to the VOTable form, which is streamed row by row to a text area or output file. The writer thus does not create a tree structure in memory. The memory requirement is therefore substantially reduced and very large VOTables can be written.

This version of the writer is a Java standalone application and provides the VOTable data only in pure XML format and not in the VOTable FITS and binary formats.

VOTFilter

VOTFilter (http://services.china-vo.org/vofilter/) is an XML filter for OpenOffice Calc developed by Chinese Virtual Observatory (China-VO) consortium. Using the filter, you can read/write files in VOTable format, edit and analyze VOTable data using OpenOffice Calc. It is a fully integrated package tool.

Here is a list of its main features:

- Fully integrated with OpenOffice;
- Reads VOTable data as other native OO Calc native format files;
- Modifies VOTable data and parameters easily;
- Save as both VOTable and other Calc supporting formats.

FITS Manager

FITS Manager (http://vo.iucaa.ernet.in/~voi/fitsm.htm) is a web-based tool for viewing, creating and editing FITS files, and for converting FITS images to other image formats. The tool was developed by Ms. Pallavi Kulkarni at IUCAA as a part of the VO-India project.

The tool currently supports following features:

- Viewing FITS files;
- Creating FITS files from image and tabular data;
- Adding extensions to FITS files;
- Converting FITS files to other formats.

FITS Manager is currently available in two forms:

- Local User Version (not available yet);
- Remote User Version (http://vo.iucaa.ernet.in:8080/examples/jsp/fitsm/welcome.jsp).

SWrap

SWrap (http://terapix.iap.fr/rubrique.php?id_rubrique=49) is a program that resamples and co-adds together FITS images using any arbitrary astrometric projection defined in the WCS standard.

11.3.2 Catalogs creation and support

SExtractor

SExtractor (http://terapix.iap.fr/rubrique.php?id_rubrique=91/index.html) is a Unix program that builds a catalogue of objects from an astronomical image (FITS format). Although it is particularly oriented towards reduction of large scale galaxy-survey data, it performs rather well on moderately crowded star fields.

ACE – Astronomical Catalogue Extractor

ACE (http://wiki.astrogrid.org/bin/view/Astrogrid/AVODemo) is a Web service wrapper to the SExtractor source extraction code. ACE is planned as a part of the AVO Demonstrator that is a set of tools being developed by the AVO Team (predominantly ESO and CDS) for processing and analysing GOODS data.

11.3.3 Data center catalog warehousing

Vizier

VizieR (http://vizier.u-strasbg.fr/) provides access to the most complete library of published astronomical catalogues and data tables available on line, organized in a self-documented database. Query tools allow the user to select relevant data tables and to extract and format records matching given criteria. Can be accessed as a Web interface or Web Service.

VizieR Proxy

The VizieR Proxy allows to access VizieR catalogs within a workflow. It is a CEA web service that converts an ADQL query to the VizieR equivalent and triggers its execution. The results are returned from VizieR to the proxy which stores them in MySpace and communicates to the Job Execution System (JES) that it has completed.

11.3.4 Robotic telescopes

Robotic telescope support

eSTAR project

The eSTAR Project (http://www.estar.org.uk/) attempts intelligent agent technologies to carry out resource discovery, submit observation requests and analyze the reduced data returned from a network of robotic telescopes.

11.3.5 Inclusion of data resources (data centers) in VO

Metadata integration

UIUC OAI Metadata Harvesting software

The UIUC OAI Metadata Harvesting Project software (http://uilib-oai.sourceforge.net/) for creating OAI Providers & Harvesters is implemented in Visual Basic and Java, and includes various stand-alone packages, plus object libraries which can be used to develop custom Providers or Harvester.

Creation of the VO nodes for the integrated search

11.3.6 Discovery of resources (services)

Metadata-based discovery in registries

AstroGrid Registry

AstroGrid Registry (http://www.astrogrid.org/maven/docs/HEAD/registry/multiproject/astrogrid-registry/) uses an eXist XML database for storing and

querying the resources of the Registry. The OAI toolkit used is the OAICat toolkit.

11.3.7 Support of collaboratory dataspaces (MySpace)

Management of dataspaces

MySpace support in AstroGrid

MySpace (http://www.euro-vo.org/twiki/bin/view/Avo/MySpace) is a component for storing user's own data within the VO. It provides a unified, virtual file-system pointing to storage in many locations. MySpaceService (http://wiki.astrogrid.org/bin/view/Astrogrid/ MySpaceService), MySpaceDirectoryService (http://wiki.astrogrid.org/bin/view/Astrogrid/ MySpaceDirectoryService), MySpaceManager (http://wiki.astrogrid.org/bin/view/ Astrogrid/MySpaceManager), MySpaceClient (http://wiki.astrogrid.org/bin/view/ Astrogrid/MySpaceClient), MySpaceCLI (http://wiki.astrogrid.org/bin/view/Astrogrid/ MySpaceCLI) are provided.

11.3.8 Integrated access to catalogs

CDS Plugin for Cross-match in the AVO prototype

The cross-match tool (http://www.euro-vo.org/twiki/bin/view/Avo/CDSXMatchPlugin) exists in 2 versions:

- a built-in function in the AVO prototype (Aladin);
- a remote HTTP service.

The plugin takes as an input 2 catalogue planes (as VOTable), and the columns containing RA and DEC in each plane:

- plane A contains the reference objects, i.e. objects for which we seek counterparts (N_A objects);
- plane B, in which counterparts are to be found (N_B objects).

Objects are considered matching if they are separated by a distance d between two thresholds: $d\_min <= d <= d\_max$ (i.e. search within rings centered on sources from list A) (distances are in arcsec). Setting d_min=0 is useful to find all matches within some radius d_max. The plugin provides on the output a new VOTable file, containing on each line:

- all the fields from plane A;
- all fields from plane B;
- the distance between t. 2 sources (in arcsec)

11.3.9 VO data processing and analysis

Type specific data analysis

Specview

Specview (http://www.stsci.edu/resources/software_hardware/specview) is a Java application for 1-D spectral visualization and analysis of astronomical spectrograms. It is capable of reading all the Hubble Space Telescope spectral data formats, as well as data from a few other instruments (such as IUE, FUSE, ISO, FORS and SDSS), preview spectra from the STScI archive, and data from generic FITS and ASCII tables. Currently supported formats: various instruments and modes (not all) of HST, IUE, FUSE, SDSS, ISO, VLT, 2dF as well as generic 2D FITS tables and text based input

Montage

Montage (http://montage.ipac.caltech.edu/) is a software system for generating astronomical image mosaics according to user-specified size, rotation, WCS-compliant projection and coordinate system, with background modeling and rectification capabilities. This release of Montage can be run on a single processor, a cluster, a computational grid or a supercomputer.

Montage processes input images that comply with the FITS standard, including specification of the WCS-compliant projection. The output image will be compliant with the FITS format standard. Users input the specification of the output image by editing an ASCII file that contains the FITS header keywords; these keywords will be written into the header of the output mosaic.

Current release supports:

- Computation of mosaics on computational grids, clusters and single processor machines. It supports two instances of parallel computing technology:
  - Message Passing Interface (MPI);
  - Planning for Execution in Grids (Pegasus), developed at the Information Sciences Institute (ISI), University of Southern California, in support of the GriPhyN project.
- Improvements to the computational algorithms:
  - Fast reprojection between tangent-plane projections;
  - Co-addition of arbitrarily large files.

11.3.10 Object kind specific data analysis

PEGASE

PEGASE (http://www2.iap.fr/pegase/) is a code which computes the spectral evolution of galaxies. The evolution of the stars, gas and metals is computed according to user selected star formation laws and initial stellar mass function. The stellar evolutionary tracks extend from the main sequence to the white dwarf stage. The emission of the gas in HII regions is also taken into account. The effect of extinction by dust is also modelled using a radiative transfer code.

The code is written in FORTRAN 77, and is made up of 5 programs plus a set of supporting files. Usually PEGASE is run via command line on a UNIX/LINUX system using either a shell script, or the interactive prompt to input the parameters.

Starburst99

Starburst99 (http://www.stsci.edu/science/starburst99/) is a web based software and data package designed to model spectrophotometric and related properties of star-forming galaxies.

Starburst99 can be compared to PEGASE above. Starburst99 uses Geneva group tracks, whilst PEGASE uses those from Padova. Starburst version 4.0 allows the use of fully line-blanketed WR model atmospheres and non-LTE O-star atmospheres (Smith, Norris, & Crowther, 2002, MNRAS, 337, 1309).

As for PEGASE, Starburst99 follows theoretical stellar tracks from the zero-age main sequence (ZAMS) to their final stages. These stages include the asymptotic giant branch (AGB) and post-AGB phases for intermediate-mass stars.

GALAXEV

GALAXEV (http://www.cida.ve/~bruzual/bc2003) computes the spectral evolution of stellar populations for a wide range of star formation history and stellar population parameters. GALAXEV is a library of evolutionary stellar population synthesis models (written in Java) computed using the new isochrone synthesis code of Bruzual & Charlot (2003). This code allows one to compute the spectral evolution of stellar populations in wide ranges of ages and metallicities at a resolution of 3 Å across the whole wavelength range from 3200 Å to 9500 Å, and at lower resolution outside this range.

11.3.11 Development of theoretical (simulation) models

Oracle data mining

Two basic classes of data mining models are provided [ORADM] – predictive and descriptive. According to the predictive models, one of the observational features is chosen as the target. The model provides a way of calculating the target as a function of the rest of the features: Y=F(X1, … , Xn). Two approaches are provided: classification (predicts a class to which an object may belong with a certain probability) and regression (predicts a value of the target). According to the descriptive models, the approaches provided include: clusterization applying certain criteria of similarity (in contrast with classification, features and classes of partitioning are unknown) or associative model (looking for stable associations).

Among predictive algorithms are classification (Naïve Bayes, Adaptive Bayes, Support Vector Machines (SVM), regression, searching for essential attributes). Among descriptive algorithms are enhanced K-means, O-clustering, association search.

Since Oracle 9i, data mining is incorporated directly into Oracle DB. Algorithms are implemented as stored procedures. Parallel computations are used if possible. Specific repository contains information on models, their applications, results. Specific data mining graphical client is provided. Java API and PL/SQL – two kinds of interfaces. JDM – new standard under development.

Weka

An exciting and potentially far-reaching development in computer science is the invention and application of methods of machine learning. These enable a computer program to automatically analyse a large body of data and decide what information is most relevant. This crystallised information can then be used to automatically make predictions or to help people make decisions faster and more accurately.

Weka (http://www.cs.waikato.ac.nz/~ml/) is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Weka is open source software issued under the GNU General Public License.

Weka has its own data format, but VOTable was readily converted to that format using a XSL stylesheet, which is available (http://www.euro-vo.org/pub/articles/ScienceWithProtoVOtools/vot2weka.xml) and which was

created from the stylesheet for VOTable-to-Mirage transformation via a few modifications.

CViz

CViz (http://www.alphaworks.ibm.com/formula/CViz) is a Java-based visualization tool designed for analyzing high-dimensional data (data with many elements) in large, complex data sets. CViz easily loads the data sets, displays the most important factors relating clusters of records, and provides full-motion visualization of the inherent data clusters.

This tool is very useful for analysts who often use statistical methods. It helps them decide on the best factors to use in multiple regression analysis and other statistical methods, primarily because it is based on linear discriminant analysis and it conforms to rigorous standards in statistical mathematics.

CViz, which is a tool for visualizing clusters in data. It reads in data in CSV format, with the column names in the first row of the file: we provide the XSL stylesheet to convert from VOTable to CViz format and the resulting CSV file. CViz is launched by invoking java on its .jar file, with additional command line options to set memory allocation, attach the HTML manual, etc.

11.3.12 User access to VO

Portals

AstroGrid Portal

The portal (http://www.astrogrid.org/maven/docs/HEAD/portal) has been designed as a framework into which components, such as workflow, can be deployed. The portal component will provide GUI pages for all components of the AstroGrid infrastructure which require them and a mechanism for third party tools providers to describe the pages they require so that users may interact with their services.

Atlasmaker

The Atlasmaker project (http://www.cacr.caltech.edu/projects/nvo/atlasmaker/) is using Grid technology in combination with NVO interoperability to create new knowledge resources in astronomy. The product is a multi-faceted, multi-dimensional, scientifically trusted image atlas of the sky, made by federating many different surveys at different wavelengths, times, resolutions, polarizations, etc. Atlasmaker does resampling and mosaicking of image collections, and is well-suited to operate with the Hyperatlas standard. Requests can be satisfied via on-demand computations or by accessing a data cache. Computed data is stored in a distributed virtual file system, such as the SRB (Storage Resource Broker).

These atlases are expected to be a new and powerful paradigm for knowledge extraction in astronomy, as well as a way to build educational resources. The system is being incorporated into the data analysis pipeline of the Palomar-Quest synoptic survey, and is being used to generate all-sky atlases from the 2MASS, SDSS, and DPOSS surveys for joint object detection.

Aladin

The *Aladin interactive atlas* is available in three modes: a simple previewer, a Java applet interface and a Java Standalone interface. Aladin (http://aladin.u-strasbg.fr/aladin.gm) is an interactive software sky atlas allowing the user to visualize digitized images of any part of the sky, to superimpose entries from astronomical catalogs or personal user data files, and to interactively access related data and information from the SIMBAD, NED, VizieR, or other archives for all known objects in the field.

Aladin is particularly useful for multi-spectral cross-identifications of astronomical sources, observation preparation and quality control of new data sets (by comparison with standard catalogues covering the same region of the sky).

The CDS provides the images of the Two Micron All Sky Survey (2MASS) from the University of Massachusetts and IPAC (JPL/ Caltech), the images of the Space Telescopes Science Institute Digital Sky Survey (DSS-I, DSS-II), with complete sky coverage, as well as an ensemble of higher resolution images (ESO-R and SERC plates) digitized at the MAMA facility in Paris. These latter images are intended for studies of crowded regions of the southern sky, such as the southern Galactic Plane. Additionally Aladin Java allows one to access other archive images such as HST, SUPERCosmos, FIRST, NVSS, Merlin, XMM-Newton, Chandra, and all images provided by SkyView (HEASARC).

Aladin can collaborate with external java tools called in this context "Aladin plugins". The first public plugin is VOPlot 2-D drawer by VO-India.

Aladin can be used as an astronomical portal displaying images and data coming not only from the default Aladin servers but also from any astronomical http servers (by the script parameter).

VO Enabled Mirage

Mirage (http://cm.bell-labs.com/who/tkh/mirage/index.html) is a Java-based software tool for exploratory analysis and visualization of images and multi-dimensional numerical data from an arbitrary domain of study. The tool shows projected images of points, point classes, or proximity structures in one, two, or higher dimensional subspaces, in linked views of tables, histograms, scatter plots, parallel coordinate plots, graphs, and trees, and over image or hypertext backgrounds linked with the data. It also provides facilities for

arbitrary plot configuration, manual or automatic classification, and intuitive graphical querying. Analysis and visualization operations are controlled by a small, interpreted command language.

VO Enabled Mirage is a wrapper around Mirage, without modification of Mirage itself.

IVOA Client package

The IVOA Client package (http://skyservice.pha.jhu.edu/develop/vo/ivoa/) is a set of classes (Java Class Library) which can be used by any Java application to retrieve VO data. It is currently used to provide VOTable functionality in VO Enabled Mirage. VOTable parsing is based on JAVOT and SAVOT.

The following public classes are included in the IVOA Client package:

- DownloadTask – a concrete Task implementation used for downloading data, which is a very common task with this package;
- ErrorPrompter – convenience class for notifying the user about exceptions or errors that may occur using graphical prompts;
- NameURLPair – data-bearing class which associates URLs returned by WSRetrievalChooser with a concisely descriptive name suitable to be used as a filename prefix;
- Task – an abstract class for encapsulating the notion of a task, which when used in conjuntion with TaskManager keeps the user updated on task status and which gives the user the option of cancelling tasks;
- TaskManager – a graphical component which displays task information and gives the user the ability to cancel tasks;
- VOTWrap – a class which contains a bunch of static interfaces and a static factory method for generating a generalized interface to VOTable data, wrapped around either JAVOT or SAVOT as the underlying implementation. It is basically a way to be able to interchange JAVOT and SAVOT without having to change application code;
- WSRetrievalChooser – a graphical component which presents the user with a simple interface for specifying parameters for Cone and SDSS CAS searches.

VODownload application uses the IVOA Client package to allow the user to download VO data to files on the local filesystem. Currently the tool supports SDSS CAS and VO Cone and SIAP searches. Service information is retrieved from an NVO registry.

TOPCAT (Tool for OPerations on Catalogues And Tables)

TOPCAT (http://www.star.bris.ac.uk/~mbt/topcat/) is an interactive graphical viewer and editor for tabular data (Java standalone application). It has been designed for use with astronomical tables such as object catalogues, but is not restricted to astronomical applications. It understands a number of different astronomically important formats (including FITS and VOTable) and more formats can be added.

It offers a variety of ways to view and analyse tables, including a browser for the cell data themselves, viewers for information about table and column metadata, and facilities for plotting, calculating statistics and joining tables using flexible matching algorithms. Using a powerful and extensible Java-based expression language, new columns can be defined and row subsets selected for separate analysis. Table data and metadata can be edited and the resulting modified table can be written out in a wide range of output formats.

VOPlot

VOPlot (http://vo.iucaa.ernet.in/~voi/voplot.htm) is a tool for visualizing astronomical data. VOPlot is developed in Java, and acts on data available in the VOTable format. VOPlot is available as a standalone version, which is to be installed on the user's machine, or as a web-based version fully integrated with the VizieR database.

VOPlot uses Ptplot 5.2, a 2D data plotter and histogram tool implemented in Java. Ptplot has been developed at Electrical Engineering & Computer Science department at the University of California, Berkeley and is available in the public domain. VOPlot is developed as an applet, and a Java plugin is required to view it in a browser.

VOSpec

VOSpec (http://pma.standby.vilspa.esa.es:8080/vospec/index.html) is an ESAC Java-based Tool for VO Spectra Handling (Spectral plotting, SED visualization).

11.3.13 Development of digital libraries for the educational resources in astronomy

DLAE metadata registries

DSpace

DSpace (http://www.dspace.org/) is a digital library system to capture, store, index, preserve, and redistribute the intellectual output of a university's research faculty in digital formats. Developed jointly by MIT Libraries and Hewlett-Packard (HP), DSpace is now freely available to research institutions world-wide as an open source system that can be customized and extended.

DSpace accepts any type of digital content, including: Text, Images, Audio, Video; documents such as articles, preprints, working papers, technical reports, conference papers, Books, Theses, Data sets, Computer programs, Visual simulations and models. Multiple formats of the same content item can be submitted to DSpace, for example, a TIFF file and a GIF file of the same image.

The DSpace submission process allows for the description of each item using a qualified version of the Dublin Core metadata schema. These descriptions are entered into a relational database, which is used by the search engine to retrieve items.

Fedora

Fedora (http://www.fedora.info/) is an open-source digital object repository management system based on the Flexible Extensible Digital Object and Repository Architecture (Fedora). The system demonstrates how distributed digital library architecture can be deployed using web-based technologies, including XML and Web services.

Fedora digital objects are encoded in XML using the Metadata Transmission and Encoding Standard (METS). The use of METS is beneficial since Fedora digital objects are encoded using a community-accepted standard.

Fedora automatically indexes the primary Dublin Core record for each digital object, as well as selected Fedora-specific metadata fields. This metadata is accessible using the OAI Protocol for Metadata Harvesting, v2.0.

UIUC OAI Metadata Harvesting software

The UIUC OAI Metadata Harvesting Project software (http://uilib-oai.sourceforge.net/) for creating OAI Providers & Harvesters is implemented in Visual Basic and Java, and includes various stand-alone packages, plus object libraries which can be used to develop custom Providers or Harvester.

It includes:

- ASP OAI 2.0 Data Provider for Database;
- ASP OAI 2.0 Data Provider for File System;
- ASP OAI 2.0 Data Provider for File/Database Hybrid;
- JSP OAI 2.0 Data Provider for Database;
- ASP OAI 2.0 Static Repository Gateway;
- Java OAI Harvester;
- VB OAI Harvester;
- VB OAI Harvester Object Lib (with a command line harvester).

11.3.14 Supervision

Data access management (authorization)

AstroGrid security

AstroGrid security (http://www.astrogrid.org/maven/docs/HEAD/security) package implements a single-sign-on (SSO) authentication system between SOAP clients and SOAP services. Messages sent with the aid of this package are annotated with SSO credentials in the SOAP headers. Messages received by services using this package are authenticated using the credentials in the message. The facade hides details of the authentication from the client and service code. Clients use a helper class, ClientSecurityGuard, to set up SSO credentials and to configure service proxies to use the credentials. Services use a matching helper class, ServiceSecurityGuard, to determine the results of authentication and to extract credentials after successful authentication. The security package ties together several Java standards:

- SOAP with Attachments API for Java (SAAJ) for handling the messages;
- JAX-RPC for the handlers that encode and parse the credentials;
- Java Authentication and Authorization System (JAAS) for authentication.

11.3.15 Workflow management

AstroGrid Workflow

This concept applies to two aspects of the system (http://wiki.astrogrid.org/bin/view/Agdoc/ArchOverview): construction and storage of a workflow; and job submission, control and notification. A portal page (or set of pages) will present the user with a list of services (retrieved from the registry). Services can be added to a workflow in sequence or parallel; the user will be prompted to provide the inputs and outputs for each service (where output from one may form the input to another, possibly with a translation service in between). The workflow can then be stored and may be retrieved later. The workflow is stored in XML form using a language derived from BPEL (formerly BPEL4WS).

A workflow can be retrieved and submitted to a job controller. This will instantiate a workflow as a series of actual tasks to be submitted to specific services and processors. The controller will be notified when a task completes (or will check if a task is overdue, in case it has crashed). Task progress information may be sent to the user according to options allowed by the workflow and by methods set up by the user in their community profile.

Pegasus

Pegasus (http://pegasus.isi.edu/) is a flexible framework that enables the mapping of complex scientific workflows onto the Grid.

11.3.16 Job control

AstroGrid Job Execution System

Job Execution Service (JES) is an AstroGrid sub-system built as a web service. It is JES that calls the web-services designated in a workflow. Job orientation *defines* AstroGrid's architecture. No useful work can be done on AstroGrid without building a workflow and submitting a job. The most-important parts of the architecture are the interfaces between the web portal and JES and between JES and the data-handling services. These interface are AstroGrid's Common Execution Architecture (CEA). Job orientation demands that the system record the specification of each job. CEA is a mechanism for recording those specifications and for passing them to web services. This descriptive power means that all CEA services have the same interface (same WSDL contract) and execution of any workflow, for any astronomical purpose, can be described by the same use case.

11.3.17 Provision of general infrastructure

Web services

Java Web Services Developer Pack

The Java Web Services Developer Pack (Java WSDP, http://java.sun.com/webservices/downloads/webservicespack.html) is a free integrated toolkit that allows Java developers to build and test XML applications, Web services, and Web applications with the latest Web services technologies and standards implementations.

Grid

Globus Toolkit

The Globus Toolkit (http://www-unix.globus.org/toolkit/) is an open source software toolkit used for building grids. It is being developed by the Globus Alliance and many others all over the world. The open source Globus Toolkit is a fundamental enabling technology for the "Grid," letting people share computing power, databases, and other tools securely online across corporate, institutional, and geographic boundaries without sacrificing local autonomy. The toolkit includes software services and libraries for resource monitoring, discovery, and management, plus security and file management.

GT3.2 implements services with a combination of C and Java. The C components run on Unix platforms, including Linux.

OGSA-DAI

OGSA-DAI (http://www.ogsadai.org.uk/) aims to provide an extension to the Open Grid Services Architecture (OGSA) specifications to allow data resources, such as databases, to be incorporated within an OGSA framework. Through the OGSA-DAI interfaces, disparate, heterogeneous data resources can be accessed and controlled as though they were a single logical resource. OGSA-DAI components also offer the potential to be used as basic primitives in the creation of sophisticated higher-level services that offer the capabilities of data federation and distributed query processing within a Virtual Organization (VO). As well as providing software, the interfaces being developed and implemented will be standardized through the Global Grid Forum (GGF) Database Access and Integration Services (DAIS) Working Group (WG).

OGSA-DAI Client Toolkit

Client Toolkit (http://www.ogsadai.org.uk/) is a Java API providing the basic building blocks for OGSA-DAI client development. Using these building blocks, a developer can construct anything from a basic query client to a complex distributed data integration client with relative simplicity. The Client Toolkit minimizes the specialist knowledge required to interact with OGSA-DAI services and shields the developer from future changes.

The main concept of the Client Toolkit is that of an *activity*. An activity dictates an action to be performed by a Grid Data Service. OGSA-DAI provides many different types of activity to perform operations such as SQL queries, XSL transformations and FTP data delivery. A sequence of one or more activities can be chained together to form a request. The Client Toolkit provides a simple mechanism for constructing and processing requests using OGSA-DAI services. The steps involved in a typical interaction are summarized below:

- Locate a Grid Data Service Factory;
- Use it to create a Grid Data Service;
- Construct a number of activities;
- Chain them together to form a request;
- Process the request using the Grid Data Service;
- Process the results of the activities;
- Terminate the Grid Data Service.

Available VO Services

XML Web Services are essentially library modules or API that live on the Web. The published methods can be accessed by making SOAP requests. Because of the underlying technology (XML, SOAP, WSDL, etc.) Web Services are inherently interoperable. They can be used regardless of what the favorite platform and operating system are. Freely downloadable toolkits, such as Microsoft's .Net Framework and Apache's Axis for Java, make the integration of Web Services seamless with the existing code. Below is a rapidly growing suite of services developed for the IVOA:

- Virtual Observatory Services;
  - Spectrum Services;
  - Filter Profiles;
  - Searchable VO Registry;
  - NED Web Services;
  - Cosmological Distance Calculator;
  - ADQL Translator;
- SDSS Related Services;
  - Image Cutout;
  - Catalog Archive Services;
- CDS Services;
  - Astronomical Coordinates;
  - VizieR Catalogues;
  - UCD Lookup.

11.4 Conceptual infrastructure of an RVO prototype

Here an attempt to produce a concretization of the layered conceptual infrastructure of RVO is undertaken. This concretization has been done assuming that initially two Data Centers will be formed – one by SAO and another one by INASAN. SAO Data Center Infrastructure is given on Fig. 14. INASAN Data Center Infrastructure is given on Fig. 15. The structures shown do not pretend to be complete or final. SAO and INASAN resources were selected for illustrative purposes. Data analysis facilities for Resource and Data Center layers were selected mostly from the facilities available off-the-shelf.

RVO Infrastructure based on Data Centers of SAO and INASAN is shown on Fig. 16. Other Data Centers can be added. This infrastructure provides an access to the Russian resources as well as to the international VO resources.

SAO DC (Fig. 14) is planned to be based on the following resources:

- Catalog of the RATAN radio sources;

- "Galaxy evolution" information systems;
- RATAN-600 archive of observation;
- 6m telescope archive of observation;
- CATS system ("warehouse") for support of astronomical catalogs.

To include RC-catalog into the VO, it is required to publish RC-catalog as a Sky Node on the Resource layer. RC-catalog is accessible on the Resource layer by means of the Web interface of CATS.

To include archives of observations into VO, the following minimal set of DAL services should be developed:

- Simple Image Access;
- Simple Spectral Access;
- ConeSearch.

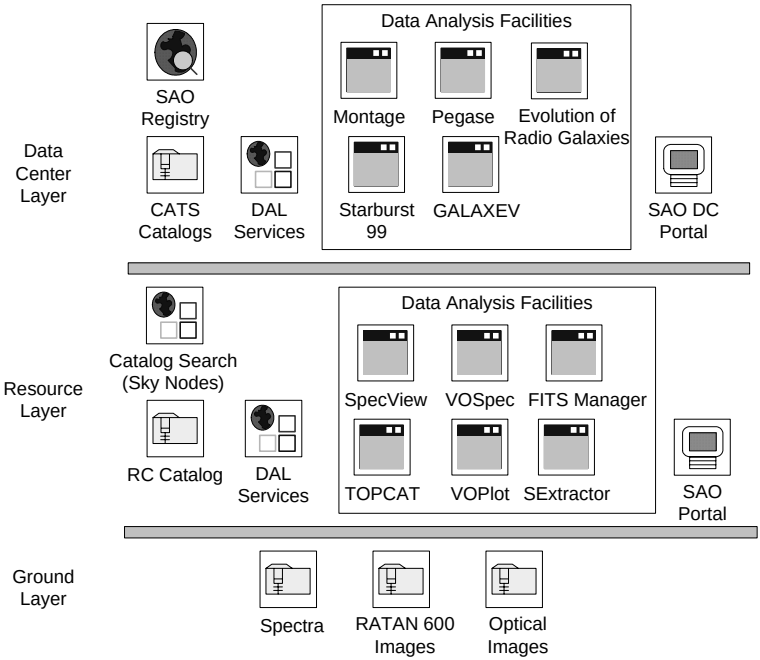

Figure 14. SAO Data Center Infrastructure

SAO DC metadata registry should conform to the IVOA standards for the VO Resource Metadata. The registry should support publishing as well as searchable services implemented applying Web service technology.

To form catalogs of the archives it is suggested to develop a specific procedure and, if possible, to use the existing facilities (such as, SExtractor).

On the Resource Layer the following facilities are suggested:

- VOPlot for VOTable format data visualization;
- TOPCAT for representing table data from catalogs;
- FITS Manager for viewing, creating and editing FITS files, and for converting FITS images to other image formats;
- VOSpec for visualization of spectra;
- Specview for visualization and analysis of astronomical spectrograms.

For access to the obsolete data formats not supported by the IVOA standards other facilities can be sued (such as conVOT for mapping of data in ASCII into the VOTable format).

On the SAO DC Layer usage of astronomical object-oriented data analysis facilities are suggested: Montage, Pegase, GALAXEV, Starburst 99.

SAO DC portal should support an interface for searching resources using the metadata registry, an interface for search in the catalog warehouse CATS, an interface for access archives of observations (images and spectra) applying DAL services.

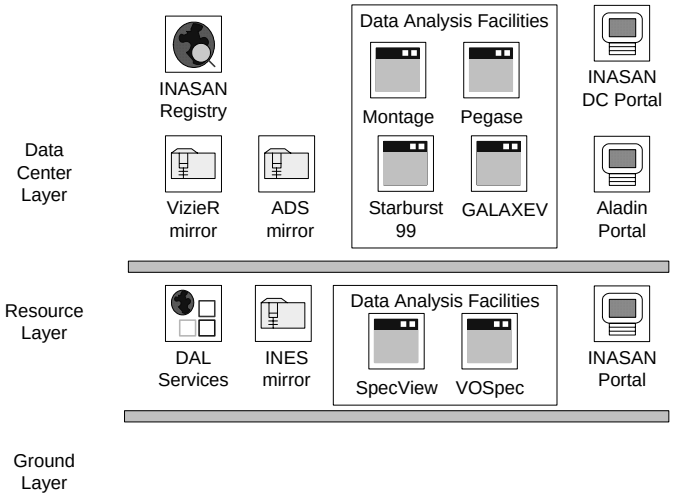

Figure 15. INASAN Data Center Infrastructure

INASAN DC (Fig. 15) is based on the following resources:

- VizieR-mirror of the data base of astronomical catalogs and data tables;
- INES-mirror of the archive data;
- ADS-mirror of the largest digital library on publications in astronomy.

Web interfaces providing access to the resources listed constitute INASAN Portal at the Resource layer.

INASAN DC metadata registry should conform to the IVOA standards for the VO Resource Metadata. The registry should support publishing as well as searchable services implemented applying Web service technology.

Data analysis facilities suggested for the Resource Layer (SpecView, VOSpec) are oriented on the INES mirror entities. Data analysis facilities suggested for the INASAN DC layer include Montage, Pegase, GALAXEV, Starburst 99.

INASAN DC portal should support an interface for searching resources using the metadata registry, an interface for search in the catalog warehouse (VizieR), an interface for access INES archives of observations applying DAL services, ADS digital library interface.

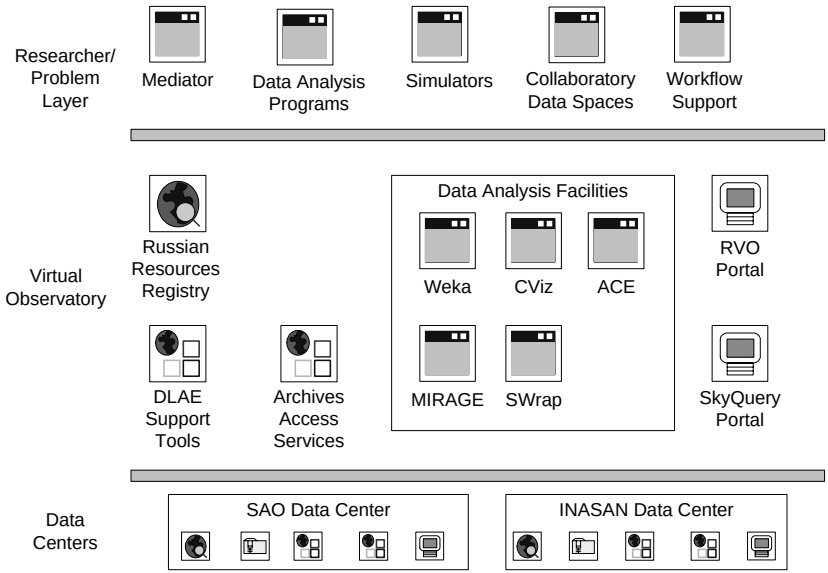

Figure 16. RVO Infrastructure

In the RVO infrastructure (Fig. 16) the RVO metadata registry should conform to the IVOA standards for the VO Resource Metadata. The registry should support searchable services implemented applying Web service technology. The registry should be accessible for the international VO for metadata harvesting.

For support of DLAE, usage of one of existing DL open sources (such as Fedora or DSpace) is planned.

For the integrated access to the catalogs of RVO the portal similar to SkyQuery is applied.

For data analysis on the VO layer the following facilities are suggested: VO enabled MIRAGE, Weka, CViz, ACE, SWrap.

IVOA Client Package can be used to provide data analysis program access from the Researcher/Problem Layer (e.g., to get data in the VOTable format).

RVO Portal provides Web-based access to various astronomical resources, including such functions as search of resources with the metadata registry, integrated search in the catalogs of sources, access to data archives (images and spectra), analysis and processing of data obtained, data visualization, access to DLAE holdings.

For one of the first versions of the RVO portal the existing client of AVO demo 2005 (http://www.euro-vo.org/twiki/bin/view/Avo/AvoDemo2005) can be used as a core and its extensions with plugins for not available functions (e.g., for integrated access in catalogs, or for access to DLAE holdings) can be undertaken.

12 Work packages for development of the RVO infrastructure

The RVO information infrastructure development program is considered as a tentative proposal open for further discussions and refinements. Formulating the work packages for the program, it was taken into account that Russia starts work on the RVO infrastructure with a several years delay comparing to other most advanced VO development projects (such as NVO, AstroGrid, AVO). The RVO program should rely as much as possible on the experience and results obtained and planned in the advanced VO projects. At the same time, the RVO project should take into account peculiar features of the state of astronomy, the respective research community and instruments in Russia. Another important consideration is that architectural and technological issues in the program are inseparable of the basic organizational steps required to implement the program. Some of them are reflected in the proposed work packages.

Due to the fact that Russia is a European country, it is assumed that the RVO information infrastructure project will be coordinated with the Euro-VO project that has started recently and is planned to be completed in 2008. Euro-VO has an intention to provide full internationalization of the Euro-VO program, designing customization tools for deployment across Europe, and mix-and-match integration with other projects. Technically RVO is planned as a hierarchy of Web services (later as a set of GRID nodes) interoperable with Euro-VO. Coordination with Euro-VO will enable the Russian scientific community to access the potential capabilities of the International VO at an early stage of its implementation; drive the evolution and implementation of the RVO by actual and specific research objectives which will complement the more general scientific requirements; educate and train scientific communities

in Russia, develop tutorials and training courses that illustrate the scientific potential of the RVO and explain in detail how to formulate and solve scientific problems using RVO. The EURO-VO technical infrastructure (after completion) is to be deployed and to be uptaken by data providers in Russia. It is required to form the vision of this process for the RVO infrastructure development and for keeping archive providers aware of this initiative. Methods and tools developed will be tested during the project for data intensive problems formulation in the context of virtual observatories. Respective research problems, subject domains and mediators are to be defined, information resources are to be selected and registered.

Specific attention requires establishing agreement on the policy for the publication of the Russian astronomical journals to join collaboration between the European journal, *Astronomy & Astrophysics* and observatory archives as an important factor of international integration in presence of VO.

1. Develop and deploy of the RVO information infrastructure;
   a) Establish the RVO WG on the information infrastructure;
   b) Develop of standards and protocols, and provide for their international acceptance;
      - Participate in the IVOA WGs for the IVOA standards development;
      - Assess IVOA standards and study the feasibility of their incorporation in RVO;
      - Decide what new standards are required for RVO and define those standards with international partners in IVOA;
   c) Define the RVO infrastructure and create designs of infrastructure components;
      - Make RVO project a member of transnational activities of Euro-VO (including participation at the technical planning meetings of Euro-VO);
      - Analyze existing VO software products and prototypes to be used in RVO. Identify the tools for re-use;
      - Establish an agreement (coordination plan) with Euro-VO regarding re-use of relevant VO software components in RVO;
      - Deploy existing components (such as DAL and generalized cross-correlation services for distributed catalogs with web and grid service support, OpenSkyNode services to provide open database access in the international VO);

- Deploy registries support (resource metadata, publishing and harvesting protocols, query protocols) to use registry services for publication, discovery, and utilization of VO resources;
- Deploy (develop) data access management (e.g., authorization) facilities;

d) Create designs of new infrastructural components of RVO;

- Develop trial versions of infrastructure components, tools, and services;
- Provide a final architecture design for RVO;
- Develop the RVO trial version;

e) Plan evolvability of the RVO infrastructure in step with continuing evolution of underlying IT, including grid technology and digital libraries (e.g., strategy for evolving from web services to the grid technology (based on OGSA DAI, WSRF));

- Establish contacts with GGF and OGSA DAI;
- Join EGEE;

f) Establish infrastructural research collaborations (possibly with the Euro-VO VOTECH project);

- Investigate feasibility of the mediation approach for the astronomical problem definition and solving;
- Investigate ontological approach for semantic definition and interoperability of astronomical resources and for their registration at mediators, expanding of metadata registries to concept spaces;
- Develop trial version of the mediation facilities;

2. Uptake of the RVO infrastructure by data providers (resources and data centers development and inclusion into the International VO);

a) Establish the RVO data provider (data center) WG;

b) Create interoperable resources;

- Provide nodes (similar to SkyNodes today) for each of the resource making them interoperable with other International VO resources;
- Develop DAL access services for resources;
- Develop catalogs;
- Provide single catalog search service;
- Establish publishing metadata registries;
- Establish user support (portal) services;

c) Create Data Centers;
   - Participate in Euro-VO Data Centre Alliance (DCA), DCA Network Workshop, DCA Committee membership (include national representatives from Russia);
   - Establish and maintain searchable resource metadata registries;
   - Develop catalog warehouses;
   - Provide integrated catalog search service;
   - Establish user support (portal) services;

3. Support of the research and development community to utilize the RVO infrastructure and data content to discover new knowledge and build new capabilities;
   a) Establish the RVO astronomy research WG;
   b) Identify astronomy research activities in Russia influencing the RVO infrastructure;
      - Analyze and identify research problems under investigation in Russia that require VO facilities; classify the problems, identify subject mediators required;
      - Define problem domain data models (including ontologies, space-time and regions requirements, etc.);
      - Identify theoretical simulation-based problems and plan computational experiments including comparison of observed and simulated data;
   c) Deploy collaboratory dataspace support facilities;
   d) Deploy (develop) workflow management facilities;
   e) Deploy job control facilities;
   f) Deploy the data analysis facilities;
      - Deploy type specific and astronomy object specific analysis tools;
      - Deploy data mining facilities over VO data (including image domain);
      - Deploy statistical analysis packages;
   g) Construct new tools to do science with the data (including user tools and services);
   h) Develop portals and interface specifications to allow research projects to use RVO;
   i) Develop simulation models;
      - Define and run the models;

- Analyze simulation results;
- Publish simulation results;

j) Define and develop subject mediators for the identified problem classes;

k) Develop tutorials and training courses that illustrate the scientific potential of the RVO and explain in detail how to formulate and solve scientific problems using RVO;

4. Develop robotic telescopes;
   a) Establish the RVO telescope WG;
   b) Workout the robotic telescope development program;
   c) Design robotic telescope service interfaces for RVO;
5. Develop Digital Library for Astronomy Education (DLAE);
   a) Establish the RVO education-oriented WG;
   b) Develop metadata standard(s) for the DLAE resources;
   c) Investigate a possibility of incorporation of the DLAE metadata in the IVOA resource registry;
   d) Deploy facilities for DLAE support (metadata registry, portal services, etc.);
   e) Make analysis and identify important resources for inclusion into DLAE;
   f) Develop a proposal for education-oriented access facilities to VO resources;
   g) Populate DLAE with information;
   h) Analyze educational application and courses that could benefit from VO data access.

References

[APALS] Abell, G.O. The National Geographic Society-Palomar Observatory Sky Survey. Astronomical Society of the Pacific Leaflets, Vol. 8, p.121, 1959

[AGRID] AstroGrid - http://www.astrogrid.org

[ASTVO] Astrophysical Virtual Observatory (AVO) - http://www.euro-vo.org

[BALAD] Boch, T., Fernique, P., Bonnarel, F. New features for Aladin 2.0, ADASS XIII, Proceedings of the conference held 12-15 October, 2003 in Strasbourg, France. ASP Conference Proceedings, Vol. 314. San Francisco: Astronomical Society of the Pacific, 2004, p.221

[BREGM] Briukhov D.O., Kalinichenko L.A., Skvortsov N.A. Information sources registration at a subject mediator as compositional development. In Proceedings of the Conference on Advances in Databases and Information Systems (ADBIS), Springer, LNCS, Vilnius, September 2001

[BSKYQ] Budavári, T., et al. 2004, Open SkyQuery - VO Compliant Dynamic Federation of Astronomy Archives, Proceedings of Astronomical Data Analysis Software and Systems XIII, 2004

[BFNVO] Building the Framework for the National Virtual Observatory. Annual Report. September 2003, AST0122449

[DPALI] Djorgovski,S., et al The Palomar Observatory-ST ScI Digital Sky Survey. Program Definition and Status. Bulletin of the American Astronomical Society, Vol. 24, p.750, 1992

[DLESN] The DLESE Program Center: Providing Infrastructure for a Distributed Community Library, NSF Proposal, http://www.dlese.org/documents/proposals/DPC_Infrastructure_dist2.pdf, 2002.

[DLOMN] Draft Standard for Learning Object Metadata, IEEE P1484.12/D5.0, 11 November 2000

[ESIMB] Egret, D., Genova, F., et al. The CDS information services in the VO era: a status report. Bulletin of the American Astronomical Society, Vol. 34, p.1104, 2002

[EURVO] Euro-VO - http://www.euro-vo.org

[EUINT] European Virtual Observatory Integration – VO-INT Proposal, 2003

[EUNET] European Virtual Observatory Data Centre Network - VO-NET Proposal, 2003

[EUTEC] European Virtual Observatory - VO Technology Centre, VO-TECH Proposal, 2003

[EVGRID] The Evolution of the Grid. Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK, 2004

[GGFOR] Global Grid Forum (GGF) - http://www.gridforum.org

[GLOBU] Globus - http://www.globus.org

[GSDSS] Jim Gray, Don Slutz, Alex S. Szalay et al Data Mining the SDSS SkyServer Database. Technical Report, MSR-TR-2002-01, January 2002

[IVOAL] International Virtual Observatory Alliance (IVOA) - http://www.ivoa.net

[KCSYN] Kalinichenko L.A. SYNTHESIS: the language for description, design and programming of the heterogeneous interoperable information resource environment. Institute for Problems of Informatics, Russian Academy of Sciences, Moscow, 1995

[KDLEU] Kalinichenko L.A. et al. Digital Libraries in Education. Analytical Survey. UNESCO Institute for Information Technologies in Education, Moscow, 2003

[KREWR] Kalinichenko L.A., Martynov D.O., Stupnikov S.A. Query rewriting using views in a typed mediator environment. Proceedings of the East-European Conference on “Advances in Databases and Information Systems" (ADBIS'04), Hungary, Budapest, Springer, Lecture notes in Computer Science, Vol. 3255, September 2004

[KSMDL] Kalinichenko L.A. Subject Mediation Infrastructure and Digital Libraries. Proceedings of the International Conference on Digital Libraries. New Delhi, February, 2004

[KDORC] A. I. Kopylov, Yu.N. Parijskij, N.S. Soboleva, A.V. Temirova, O.V. Verkhodanov, Distant object selection from the RC-catalog by their radio properties, in Proceedings of Russian Astronomical Conference – 2004, p. 35

[LOAIF] C. Lagoze and H. Van de Sompel, “The Open Archives Initiative: Building a low-barrier interoperability framework,” Joint Conference on Digital Libraries, Roanoke, VA, 2001.

[LSSTO] Large Synoptic Survey Telescope: Overview, J.A. Tyson and LSST Collaboration, huhepl.harvard.edu/~LSST /general/LSST_overview_tyson.pdf

[LTHEO] Gerard Lemson (GAVO), Jörg Colberg (NVO). Theory in the VO. White paper, 2004 http://www.ivoa.net/pub /papers/TheoryInTheVO.pdf

[MPALQ] A. Mahabal. Exploring the Time Domain with the Palomar-QUEST Sky Survey. astro-ph/0408035

[MSQYQ] Tanu Malik Alex S. Szalay Tamas Budavari Ani R. Thakar. SkyQuery: A Web Service Approach to Federate Databases. Proceedings of the 2003 CIDR Conference

[NVOBS] National Virtual Observatory (NVO) - http://www.us-vo.org

[OVIZI] Ochsenbein, F., Bauer, P., Marcout, J. The VizieR database of astronomical catalogues, Astronomy and Astrophysics Supplement, v.143, p.23-32, 2000

[OGDA5] OGSA-DAI ROAD MAP FOR Q3 2004 – Q3 2005, September 1, 2004

[OGDAI] Open Grid Services Architecture Data Access and Integration (OGSA-DAI) - see http://www.ogsadai.org.uk/

[ORADM] ORACLE data mining: http://otn.oracle.com/products/bi /odm/index.html

[PBTRI] Yu.N. Parijskij, W.M. Goss, A.I. Kopylov, N.S. Soboleva, A.V. Temirova, O.V. Verkhodanov, O.P. Zhelenkova. 2000. RATAN-600 - VLA - BTA-6m ("Big Trio") project: multicolour studying of distant radio galaxies. Atsron. Astrophys. Trans. V.19, No 3-4, PP.297-304, astro-ph/0005240

[RCOSS] Read,M.A., Hambly,N.C. The SuperCOSMOS Sky Surveys. ADASS X, ASP Conference Proceedings, Vol. 238. San Francisco: Astronomical Society of the Pacific, ISSN: 1080-7926, 2001, p.182

[RVODG] David De Roure, Mark A. Baker, Nicholas R. Jennings, Nigel R. Shadbolt. The Virtual Observatory as a Data Grid, Report of the workshop held at the e-Science Institute, Edinburgh on 30 June – 2 July 2003

[RSPCE] Russell, J.L. The Guide Star Selection System and the Guide Star Catalog for Space Telescope, ASTROMETRIC TECHNIQUES: IAU SYMP:109 FLORIDA P. 721, 1986

[SKYQU] SkyQuery: http://www.openskyquery.net

[SSDSS] Alexander S. Szalay, Peter Kunszt, Ani Thakar, Jim Gray, Don Slutz, Robert J. Brunner. Designing and Mining Multi-Terabyte Astronomy Archives: The Sloan Digital Sky Survey Technical Report, MS-TR-99-30, February 2000

[SQLHTM] Alex Szalay, George Fekete, Jim Gray. SQLServer2000 HTM Interface specification, July 2003, http://skyservice.pha.jhu.edu /develop/vo/ adql/htmdll_2_0.doc.

[USNO] D.G. Monet. The USNO-B Catalog, the Astronomical Journal, 125:984-993, 2003 February

[VCATS] Verkhodanov O.V., Trushkin S.A., Chernenkov V.N. CATS: a database system of astrophysical catalogues. Baltic Astronomy, 1997, V.6, No 2. P.275-278

[VRGAL] O.V. Verkhodanov et al. The software system "Evolution of radio galaxies", astro-ph/9912359

[VRAST] V.Vitkovskij, V.Chernenkov, A.Ivanov, V.Gurin, N.Kalinina, V.Komarov, S.Moiseev, A.Nazarenko, V.Shergin, O.Zhelenkova. The remote access system for the largest Russian Telescopes BTA and RATAN-600. Baltic Astronomy, 2000, v.9, N4, pp 527-531

[VRVOI] V.Vitkovskij, O.Zhelenkova, N.Kalinina, V.Chernenkov, and V.Shergin. The Russian Vurtual Observatory Project Insight. Toward an International Virtual Observatory, Proceedings of the ESO/ESA/NASA/NSF Conference held in Garching, Germany, 10-14 June 2002. Edited by P.J. Quinn, and K.M. Gorski. ESO Astrophysics Symposia. Berlin: Springer, 2004, p. 323.

[VIRTU] VirtU – The Virtual Universe comprising the Theoretical Virtual Observatory & the Theory/Observations Interface. A collaborative e-science proposal. Virtual Universe PPARC proposal. http://star-www.dur.ac.uk/~csf/virtU/virtU-final.pdf

[VODGR] The Virtual Observatory as a Data Grid, Report of the workshop held at the e-Science Institute, Edinburgh on 30 June – 2 July 2003

[WFIRS] White, Richard L., Becker, Robert H., Helfand, David J. The FIRST Radio Survey, ``Astrophysics and Algorithms: a

DIMACS Workshop on Massive Astronomical Data Sets", meeting abstract, 1998

[ZARCH] Zhelenkova O.P., Vitkovskij V.V., Plyaskina T.A., Malkova G.A., Shergin V.S.. The SAO RAS observation archive state and prospects of development, Bulletin of the Special Astrophysical Observatory (in press)

Glossary of acronyms

2dF	The Two Degree Field system (`2dF') is the AAT's most complex astronomical instrument.
2MASS	Two Micron All Sky Survey
ADAC	Astronomical Data Analysis Center (Japan)
ADQL	Astronomical Data Query Language
ADL	US Department of Defense's Advanced Distributed Learning Network
ADS	NASA Astrophysics Data System
ADT	Abstract Data Type
AGN	Active Galaxy Nucleus
ALMA	Atacama Large Millimeter Array telescope
APM	Astronomical Precise Machine
ASPID	Archive of Spectral, Photometric and Interferometric Data
ASU	Astronomical Server URL
AVO	Astrophysical Virtual Observatory
BPEL	Business Process Execution Language
CAD	Centre of Astronomical Data of the Institute of Astronomy RAS
CADC	Canadian Astronomy Data Centre
CATS	Astrophysical CATalogs support System
CCD	Charge Coupled Device
CDM	Cold Dark Matter
CDS	Centre deDonnйes astronomiques de Strasbourg (France)
CEA	Common Execution Architecture
CGI	Common Gateway Interface
Chandra	NASA Chandra X-ray Observatory
CM	collective memories
CMB	Cosmic Microwave Background

CVO	Canadian Virtual Observatory
DAI	Data Access and Integration
DAIS	Data Access and Integration Service
DAL	IVOA Data Access Layer
DAS	Data Archive Server
DCA	Data Centre Alliance
DCL	Data Center Layer
DLAE	Digital Library for Astronomy Education
DLE	Digital Libraries in Education
DLESE	Digital Libraries for Earth Science Education
DM	Data Model
DQP	Distributed Query Processing
DSS	Digitized Sky Survey
DSS1	Digitized Sky Survey 1
DSS2	Second Digitized Sky Survey
EDG	European Data Grid
EISCAT	European Incoherent SCATter Scientific Association
e-MERLIN	MERLIN/VLBI National Facility
EPO	Education and Public Outreach
ESA	European Space Agency
ESAI	Extended time series of Solar Activity Indices
ESO	European Southern Observatory
ESO/ST-ECF	ESO and Space Telescope-European Coordinating Facility
Euro-VO	European Virtual Observatory project
FIRST	Faint Images of the Radio Sky at Twenty-cm
FITS	Flexible Image Transport System
FORS	VLT FOcal Reducer/low dispersion Spectrograph
FP6	6th Framework Program
FRII	Fanaroff-Riley class II

FUSE	Far Ultraviolet Spectroscopic Explorer
GALEX	Galaxy Evolution Explorer
GAVO	German Astrophysical Virtual Observatory
GBT	Robert C. Byrd Green Bank 100m telescope
GCVS	General Catalogue of Variable Stars
GDQS	Grid Distributed Query Service
GDS	Grid Data Service
GGF	Global Grid Forum
GIS	Geo Information System
GQES	Grid Query Evaluation Service
GSC	Guide Star Catalog
GT	Globus Toolkit
HEASARC	High Energy Astrophysics Science Archive Research Center
HST	Hubble Space Telescope
HTM	Hierarchical Triangular Mesh
IAU	International Astronomical Union
INES	IUE Newly Extracted Spectra
INT-WFS	Wide Field Survey on the Isaac Newton Telescope (INT)
IPAC	Infrared Processing and Analysis Center
IRAC	Ifrared Array Camera on the Spitzer Space Telescope
IRSA	Infrared Science Archive
ISO	The Infrared Space Observatory is an European Space Agency (ESA) mission in cooperation with ISAS (Japan) and NASA (USA)
IUE	International Ultraviolet Explorer
IVO	International Virtual Observatory
IVOA	International Virtual Observatory Alliance
JAAS	Java Authentication and Authorization System
JAC	Joint Astronomy Centre

JDBC	Java Database Connectivity
JHU	Johns Hopkins University
JVO	Japanese Virtual Observatory
LHC	Large Hadron Collider
LS	Long Slit Spectrograph
LSST	Large Synoptic Survey Telescope
MAST	Multimission Archive at Space Telescope
MERLIN	Multi-Element Radio Linked Interferometer Network
METS	Metadata Transmission and Encoding Standard
MIGALE	Multiparametric virtual Instrument for GALaxy Evolution
MIPS	Multiband Imaging Photometer for the Spitzer Space Telescope
MOFS	Multi Object Fiber Spectrograph
MPE	Max Planck Institute for Extraterrestrial Physics
MPFS	Integral Field Spectrograph
MPG/ESO	The MPG (Max Planck Gesellschaft)/ESO 2.2m telescope at La Silla
MPI	Message Passing Interface
NASA	National Aeronautics and Space Administration
NOAO	National Optical Astronomy Observatories
NRAO	National Radio Astronomy Observatory
NSDL	National Science Digital Library
NSF	National Science Foundation
NTT	ESO New Technology Telescope
NVO	National Virtual Observatory
NVSS	NRAO VLA Sky Survey
OAI	Open Archives Initiative
OAI-PMH	OAI-Protocol for Metadata Harvesting
OGSA	Open Grid Services Architecture

OGSA-DAI	Open Grid Services Architecture Data Access and Integration
OGSI	Open Grid Service Infrastructure
OMP	Observation Management Project
OWL	100-m class optical and near-infrared telescope
OWL	OWL Web Ontology Language
POSS	Palomar Sky Survey
PPARC	Particle Physics and Astronomy Research Council (UK)
PQ	Palomar-Quest
PSC	Point Source Catalog
QSO	QuasiStellar Object
RC	RATAN-600 RC-catalog
RDBMS	Relational DataBase Management System
RL	Resource Layer
RPL	Research/Problem Layer
RTML	Robotic Telescope Markup Language
RVOII	Russian Virtual Observatory Information Infrastructure
RVO	Russian Virtual Observatory
SAAJ	SOAP with Attachments API for Java
SAI	Sternberg Astronomical Institute
SCSE	Shared Center of Science and Education
SDSC	San Diego Supercomputer Center
SED	Spectral Energy Distributions
SDSS	Sloan Digital Sky Survey
SIA	Simple Image Access
SIAP	Simple Image Access Protocol
SIRTF	Space Infrared Telescope Facility
SKA	Square Kilometre Array telescope
SOAP	Simple Object Access Protocol

SOHO	Solar and Heliospheric Observatory
SRB	Storage Resource Broker
SSA	Simple Spectral Access
SSAP	Simple Spectral Access Protocol
SSO	single-sign-on
SSS	SuperCosmos Sky Survey
STScI	Space Telescope Science Institute
SuperCOSMOS	SuperCOSMOS Sky Surveys (SSS)
SWG	Science Working Groups
SWIRE	SIRTF Wide-Area InfraRed Extragalactic Legacy survey
TCDL of IEEE-CS	Technical Committee on Digital Libraries of the Institute of Electrical and Electronics Engineers Computer Society
TNO	Transient Near-Earth Objects
TOI	Theory/Observations Interface
TOML	Telescope Observation Markup Language
TVO	Theoretical Virtual Observatory
UCD	Unified Content Descriptor
UDDI	Universal Description Discovery & Integration
UKIRT	United Kingdom Infrared Telescope
UKIRT WFCAM	Wide Field Infrared Camera For UKIRT
UML	Unified Markup Language
USNO	United State Naval Observatory catalog
UV	UltraViolet
VirtU	Virtual Universe
VLA	Very Large Array
VLBA	Very Long Baseline Array
VLBI	Very Long Baseline Interferometry
VLT	Very Large Telescopes
VTIE	Virtual Telescopes in Education

VO	Virtual Observatory
VOFC	EURO-VO Facility Centre
VOQL	Virtual Observatory Query Language
VOL	VO Layer
VOTC	EURO-VO Technology Centre
WCS	World Coordinate System
WFCAM	Wide Field Infrared Camera For UKIRT
WG	Work Group
WS	Web Service
WSDL	Web Services Description Language
WSRF	Web Services Resource Framework
WUN	Worldwide Universities Network
XML	Extensible Markup Language
XMM-SSC	XMM-Newton Survey Science Centre
XMM-Newton	XMM-Newton is a joint NASA-European Space Agency orbiting observatory
XSC	Extended Source Catalog
XSD	XML Schema Definition Language
XSL	Extensible Stylesheet Language
Yohkoh	Yohkoh Mission is a Japanese Solar mission with US and UK collaborators.