Russian Virtual Observatory Community Centre for Scientific Problems Solving over Multiple Distributed Information Sources♣

© Leonid Kalinichenko, Sergey Stupnikov, Alexey Vovchenko, Victor Zakharov

Institute of Informatics Problems, Russian Academy of Sciences
leonidkssa@synth.ipi.ac.ru, ssa@ipi.ac.ru, vintik@gcnet.ru, vzakharov@ipiran.ru

Olga Zhelenkova

Special Astrophysical Observatory, Russian Academy of Sciences
zhe@sao.ru

Abstract

The paper considers one of the first steps of implementation of the Russian Virtual Observatory Information Infrastructure (RVOII) – organization of a Community centre at IPI RAS (Moscow) for support of scientific astronomical problem solving over distributed repositories of astronomical information. As a motivation the trends for distributed infrastructure development for e-science are presented. Information infrastructure of RVO aimed to satisfy International Virtual Observatory Alliance standards is briefly introduced. Structure of AstroGrid system considered as a core of RVOII is presented in brief. First trial of scientific use of AstroGrid RVO over distributed astronomic sources is explained in detail.

1 Introduction: the trends for distributed infrastructure development for e-science

In different branches of science an exponential growth of experimental (observational) data is observed. For instance, in astronomy the current and expected rate of data growth obtained from observatories is doubled during a period from six months to one year. This rate is larger than growth of components per integrated circuit that is doubled according to the Moore's law each 18 months. Such growth leads to enlargement of a gap between researchers and information sources and brings to a necessity of looking for new ways of problem solving over multiple distributed information collections that are accumulated in specialized centers of data and computational resources. e-Science refers to science that is enabled by the routine use of distributed computing resources by end-user scientists. Computational researchers need enabling, scalable, interoperable application software to conduct examinations of their ideas and data. For e-science a number of technological solutions and infrastructures are being developed. These solutions are oriented on the adequate organization of distributed information and elimination of the gap mentioned. In the introductory section some of such solutions will be briefly presented.

Grid technology is a natural way of designing the IT infrastructure for e-Science [17]. 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 distinguished [17]: *computational grid* supporting utility computing or computing-on-demand, *information grid* involving integration of large scale distributed data repositories, *hybrid grid* combining information and computational grids emphasizing integration of experimental data and simulations, *semantic grid* where services are supported by metainformation that are used for proper service discovery and composition for specific applications.

gLite [9] is a hybrid grid middleware recently released by EGEE [8], the largest Grid Infrastructure Project currently being funded in Europe. The role of gLite is to hide the heterogeneous nature of both the *computing element*s (CEs), i.e. services representing a computing resource, and *storage elements* (SEs), i.e. services accessing the data in the files.

Technology for shared collections support distributed across multiple organizations and heterogeneous storage systems. SRB – *The SDSC Storage Resource Broker* [16]– supports shared collections that can be distributed across multiple organizations and heterogeneous storage systems. The SRB can be used as a Data Grid Management System (DGMS) that provides a hierarchical logical namespace to manage the organization of data (usually files).

The SRB software infrastructure can be used to enable Distributed Logical File Systems, Distributed Digital Libraries, Distributed Persistent Archives, and

Proceedings of the 8th Russian Conference on Digital Libraries RCDL2006, Suzdal, Russia, 2006

Virtual Object Ring Buffers. The most common usage of SRB is as a Distributed Logical File System (a synergy of database system concepts and file systems concepts) that provides a powerful solution to manage multi-organizational file system namespaces.

SRB can be evaluated as a platform for ensuring preservation and, more in generally, the long term availability of the access to digital information. Widely used content repository systems, like DSpace [20] and Fedora [13] as well as DLs are presently using the SDSC Storage Resource Broker.

Common information objects models. New object models are under development (e.g., *DoMDL* [6]) with an intention to represent a wide variety of information object types with different formats, media, languages and structures. Moreover, it can represent new types of documents that have no physical counterpart, such as composite documents consisting of the slides, video and audio recordings of a lecture, a seminar or a course. It can also maintain multiple editions, versions, and manifestations of the same document, each described by one or more metadata records in different formats. Every manifestation of the digital object can be either locally stored, or retrieved from a remote server and displayed whether at run time or in its remote location.

Digital library technologies. The OAI-Protocol [12] 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 and XML. 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).

Cheshire [14] is a next-generation online information retrieval system based on international standards which supports advanced information retrieval techniques, including probabilistic, vector space and boolean retrieval and data fusion techniques allowing the combination of results from multiple search methods. The Cheshire system supports indexing and retrieval from the SRB. The Cheshire system uses an object-oriented design that operates in both single-processor and distributed computing environments. Large digital library workloads running on small clusters and the SDSC TeraGrid across different domains.

The Cheshire digital library system implements an extensible workflow which has enabled the integration of the many components within the system. The Multivalent digital object management technology [15] is used to parse documents in multiple formats. The Cheshire workflow integrates this parsing technology with an indexing and categorization system. The capability of the Multivalent model to annotate digital entities (whether manually or automatically) is critical to generating multiple ontologies which may be used to characterize knowledge relationships.

It is planned to integrate into Cheshire an inference engine which will apply rules to the set of operations that will be performed by the application or grid service. In Cheshire there is a capability to cluster together topics which may be semantically related by automating the process of association between natural language and ontologies.

The role of computational linguistics and ontological processes. "Lexical Priming" proposes a radical new theory of the lexicon, which amounts to a completely new theory of language based on how words are used in the real world. Here they are not confined to the definitions given to them in dictionaries but instead interact with other words in common patterns of use. Opposite to classical theory, the new approach reverses the roles of lexis and grammar, arguing that lexis is complexly and systematically structured and that grammar is an outcome of this lexical structure [10]. The phenomenon of "collocation", the property of language whereby two or more words seem to appear frequently in each other's company, offers a clue to the way language is really organised. Using concrete statistical evidence from a corpus of newspaper English, but also referring to travel writing and literary text, it is argued that words are "primed" for use through our experience with them, so that everything we know about a word is a product of our encounters with it. This knowledge explains how speakers of a language succeed in being fluent, creative and natural.

The semantics, or meaning of terms, within any given domain depends on associated context. Clustering of semantic terms and relationships to produce knowledge are used with the

Cheshire and Multivalent systems to support domain analysis. Need to be combined with methods of text mining techniques based on generative grammar approaches as well as on analysis of the lexis.

Examples of new integrated infrastructures. The University of California, Berkeley and the University of Liverpool are developing a Information Retrieval and Digital Library system (Cheshire3) that operates in both single-processor and Grid distributed computing environments. Cheshire3 [14] is integrated with SRB DataGrid storage system. Because each object in the SRB is a "document" that can be processed and indexed and identified by its unique SRB ID, it can be retrieved the original from storage on demand without needing to store an XML representation of it. Since SRB is also being integrated into the DSpace Digital Library Framework, an advanced indexing and search systems for DSpace and SRB installations is planned.

OpenDLibG [7]: extending OpenDLib by exploiting a gLite Grid Infrastructure. OpenDLib has been extended in order to make it able to exploit the storage and processing capability offered by a gLite Grid infrastructure. Thanks to this, OpenDLib, applying DoMDL, is able to handle a much wider class of documents than in its original version and, consequently, it can serve a larger class of application domains. In particular, OpenDLib can manage documents that require huge storage capabilities, like particular types of images, videos, and 3D objects, and also create them on-demand as the result of a

computational intensive elaboration on a dynamic set of data, although performed with a cheap investment in terms of computing resource.

OpenDLibG DL has been applied for creating a DL supporting the work of the agencies that collaboratively work at the definition of environmental conventions. By exploiting their rich information sources, ranging from raw data sets to maps and graphs archives, these agencies periodically prepare reports on the status of the environment. The so created DL provides the data, the documents, the dynamically generated reports, and any other content and services deemed as relevant with respect to the day-by-day activity of people who have to take decisions on environmental strategies.

The remaining part of the paper is devoted to the information infrastructure of the Russian Virtual Observatory (RVO) project [18, 5] and its core that is built applying the AstroGrid system [2,3]. A combination of many of the technologies mentioned are already used and planned to be used in the project.

2 RVO project objectives

Main objectives of the RVO project (http://www.inasan.rssi.ru/eng/rvo/, [18]) 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.

3 VO architecture according to the IVOA

RVO infrastructure should be based on the International Virtual Observatory Alliance (IVOA) standards [4,5]. 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 9 IVOA Working Groups focusing on Resource Registry, Data Modeling, Content Description, Data Access Layer, VOTable, VO Query Language, VO Event, Grid & Web Services, Standards & Processes (http://www.ivoa.net/twiki/bin/view/IVOA/WebHome).

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.

In the IVOA architecture, the available services are divided 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.

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

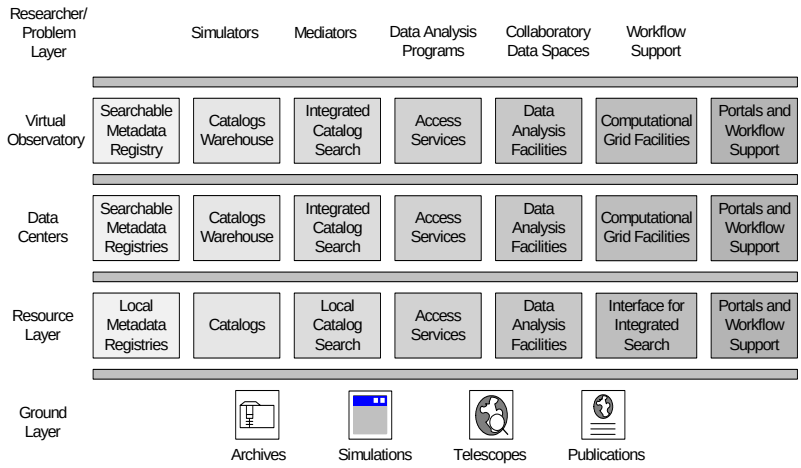

Figure 1. The RVO layered infrastructure

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. 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) [21].

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 (ontology) 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.

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 establishes 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.

4 Information infrastructure of RVO

According to the project of the RVO Information Infrastructure (RVOII) [5], the main principles of RVOII include the following:

- the architecture is represented as a network of interoperating web services (implemented in a grid architecture as soon as the required standards will mature and will be accepted by the international community);

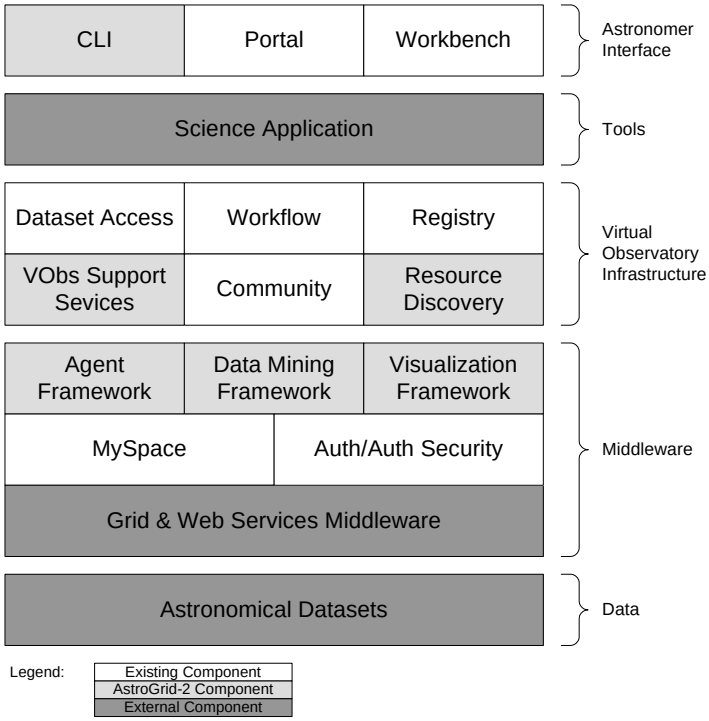

Figure 2. Existing and planned components of the AstroGrid

- moving processing to the data is another principle motivated by data intensive character of VO applications;
- 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.

Conceptually the information infrastructure of the RVO is considered to be multilayered (Fig.1). Each layer of the infrastructure [5] includes information entities, access services, data analysis and user support facilities intended for the information entities of the respective layer. Primary information sources (archives, simulation results, robotic instruments and various publications related to astronomy) form the ground layer of the 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).

Data processing and analysis facilities include 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 (workflow and portal functionalities); application program interfaces to various functions of the Resource layer are also included. 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 specialization. They should conform to common

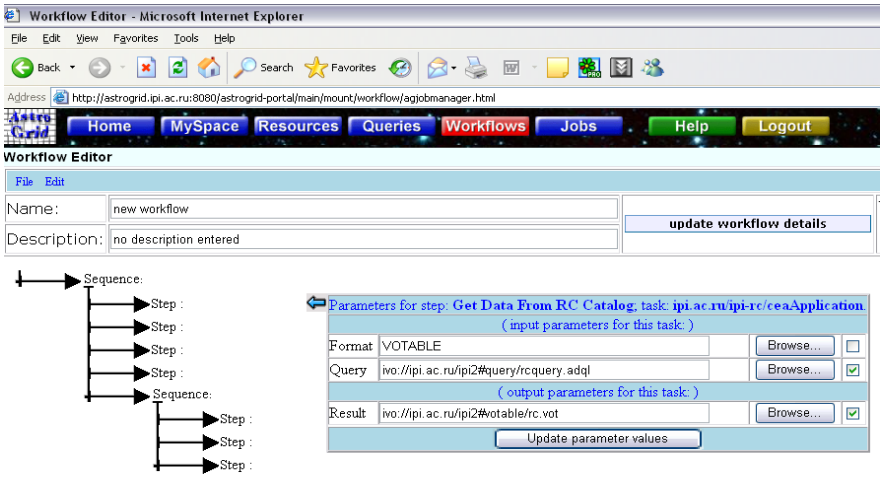

Figure 3. RC Catalogue querying

protocols and standards. Graphically the DCL components are similar to that of the RL layer.

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 the metadata harvesting according to the OAI-PMH standard [12]. 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 provide 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.

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. Important ingredients of RPL include facilities for workflow definition and management. Collaborative dataspace management is provided for user's (and user group) own data organization within the RVO.

In the RVO infrastructure (Fig. 1) 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.

5 AstroGrid RVO as the RVOII core

The AstroGrid project [2] is a part of the worldwide International Virtual Observatory (IVO) and is considered as the UK contribution into IVO. The AstroGrid project objectives include:

- creation of information and computational grid for integration of various heterogeneous sky surveys and services, accumulated in the world;
- high performance facilities for analysis of the resources mentioned and their data mining;
- instruments for interactive access to databases and their analysis;
- facilities for creation and loading programs implementing user algorithms onto servers performing data mining;
- methods of information resources discovery on the basis of the metadata registries.

AstroGrid is a part of the UK scientific program in the e-science area and is interconnected with the projects on particle physics, bio-informatics and grid technologies fundamentals. AstroGrid is also a part of the European VO (Euro-VO) project. The budget of the project for 2001 – 2004 established 8.9 millions of US dollars and for 2005 – 2007 (Astrogrid – 2 project) – 9.6 millions of US dollars. On Fig. 2 the existing components of AstroGrid are depicted as well as the components planned for the AstroGrid -2.

Registry is an OAI-PMH protocol compliant component and serves as a metadata collection defining resources that can be applied for the problem solving by the VO. Two kinds of registries are supported: publishing of information about the resources provided and harvesting for collecting metadata about resources registered at another registries of IVO.

Community is a component supporting registration and personal authentication of users.

MySpace component is a file system providing file access for all services of the AstroGrid system. MySpace is used also for storage and moving files among tasks solved by means of VO. MySpace is considered as a prototype of the VOSpace standard being developed by IVOA. This standard is planned to

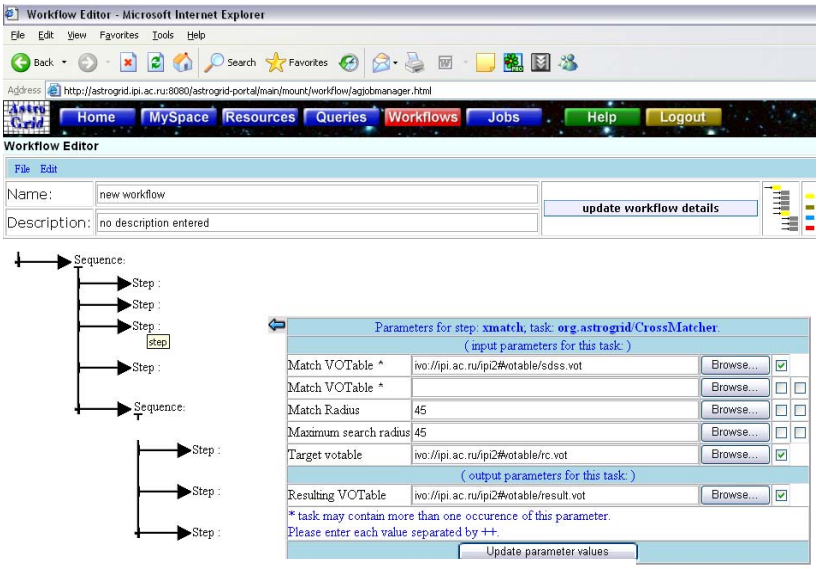

Figure 4. CrossMatch of the results obtained

allow usage of other large file systems in the grid, including SRB.

Job Execution System (JES) components is an engine for workflow execution providing for concurrent call of AstroGrid services supporting the *Common Execution Connector* (CEC) interface.

Portal and *Workbench* belong to AstroGrid client applications. Portal provides GUI making possible to manage the AstroGrid components from a browser. Another recommended client interface is Workbench that has been developed as a Java WebStart application.

Common Execution Architecture (CEA) provides for interoperability of AstroGrid services, application description, description of tasks that should be called (e.g., from a workflow), support of asynchronous activities and interactions, that preserve the state of such activities. Various details of SOAP, WSDL-contracts, asynchronous activities are hidden from the users while applying AstroGrid services in the desktop applications. Application interface *AstroGrid Client Runtime* (ACR) is used for that.

Data Set Access (DSA) component provides for creation and access to archives of data stored in the relational DBMSs (such as PostgreSQL, MySQL) using ADQL – a specific astronomical dialect of SQL, standardized by IVOA. Besides that, DSA provides an IVOA/NVO cone search interface and IVOA SkyNode interface.

An analysis shows that usage of AstroGrid as the RVOII core makes possible to implement the basic principles of RVOII (such as support of interoperable grid services, architecture modularity, possibility of re-use and composition of services, creation of the multilayer architecture). AstroGrid components are directly applicable as the core of the RVOII architecture.

6 First AstroGrid RVO trial

6.1 Distant galaxies discovery problem

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

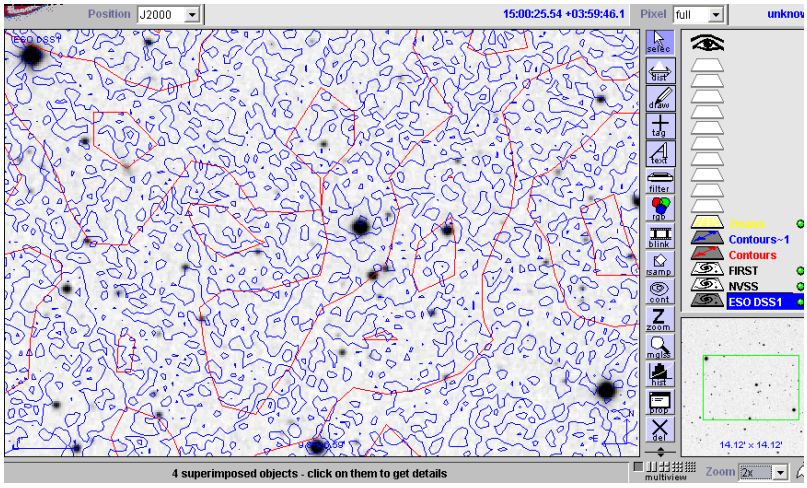

Figure 5. Example of an image obtained

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.

6.2 Information processing for distant galaxies discovery applying AstroGrid and Aladin

To process information for distant galaxies discovery applying AstroGrid special facilities has been developed, including workflow containing the required steps. The information processing is subdivided in two main phases: “Preparing the results that may contain candidates to be analyzed further” and “Result exploration by a researcher”.

On the first phase applying AstroGrid RVO the workflow is executed that selects from the RC catalogue the candidates that potentially might be classified as distant galaxies. After that for each potential candidate from the image archives the optical and radio images are retrieved and their superposition is produced. The results are stored in a format convenient for the user. On the second phase a researcher applies the Aladin Sky Atlas [1] to open the images obtained for their visual analysis.

Phase 1. The workflow defined for the AstroGrid RVO looks as follows:

- Step 1. Querying the RC catalogue;
- Step 2. Querying the SDSS DR3 [19] catalogue;
- Step 3. CrossMatch of the results obtained on the previous steps;
- Step 4. Retrieving and superposition of images.

Step 1. Querying the RC catalogue: is executed on the AstroGrid installed at IPI RAS, Moscow.

CEA application is executed to query the database containing the RC catalogue. A copy of the RC catalogue has been preliminary stored in the PostgeSQL DBMS and formed as a DSA component of AstroGrid. A result of the query in VOTable format is automatically stored in MySpace. The query is written in the ADQL language:

```
SELECT crd.ra, crd.de, cat.name
FROM RCCatalog as cat, CoordEQJ as crd
WHERE cat.coord_id = crd.coord_id
```

This step of the workflow is shown on Fig. 3. CEA application gets an ADQL query and resulting VOTable as parameters. In this specific workflow instead of the query itself an address of the file containing the query is passed as a parameter. For a result an address of the file to accept the result is passed as a parameter.

Step 2. Querying the SDSS DR3 catalogue: is executed on the SDSS server in USA.

CEA application, using web service provided by SDSS (http://voservices.net/CasService/ws_v1_0/CasService.asmx), executes the query and returns the result

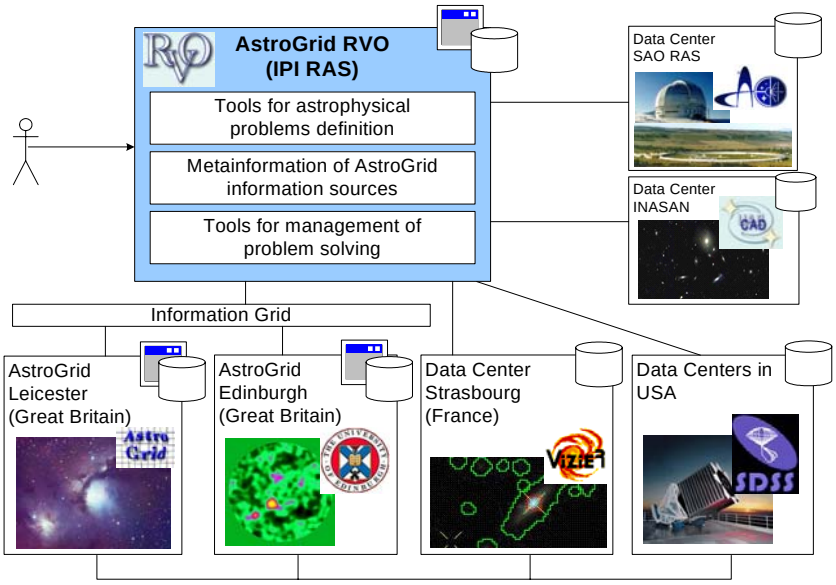

Figure 6. AstroGrid RVO as a part of the International Virtual Observatory

in the VOTable format that automatically is stored in MySpace. The SDSS query looks as follows:

```
SELECT ra=cast(ra as real),
       [dec]=cast([dec] as real),
       objid, u, g, i, r, z,
       colorIndexURG = (u+r)/2.0-g
FROM PhotoPrimary
WHERE ra BETWEEN 225.0 AND 225.5
      AND [dec] BETWEEN 4.0 AND 5.61
      AND r BETWEEN 15.0 AND 23.0
```

Step 3. CrossMatch of the results: is executed on one of the AstroGrid servers in UK.

A crossmatch (CrossMatch Full – ivo://org.astrogrid/CrossMatcher) of two VOTables obtained on the previous steps is executed. A result in the VOTable format is put into MySpace. Fig. 4 shows the definition of this workflow step in AstroGrid. Besides VOTables, Match Radius and Maximum Search Radius (chosen to be 45) are passed as parameters. Such a big radius has been chosen due to the low precision of the RC catalogue.

Step 4. Retrieving and superposition of images: is executed at IPI RAS, Moscow.

The CEA application is executed that applying Aladin, for each candidate object obtained after crossmatch (using its coordinates) retrieves optical images from the image archive DSS. Radio images from archives FIRST and NVSS are superimposed with the first image. Additionally through the SDSS DR3 and 2MASS catalogues the sources are obtained that are located in the neighborhood of the candidate object. A script written in the Aladin script language is shown below. An address variable contains coordinates in the format "RaH:RaM:RaS DecH:DecM:DecS":

```
"reset; grid;"
"get DSS.ESO(DSS1,14.1,14.1) " + adress + " 5';
sync;"
"get NVSS(0.2,15.0,Stokes I,Sine) " + adress +
" 5';"
"sync; contour 4;"
"get FIRST(10) " + adress + " 5';"
"sync; contour 4;"
"get SDSSDR3cat " + adress + " 1';"
"sync;"
"get VizieR(2mass) " + adress + " 1';"
"sync;"
"show 1; hide NVSS FIRST;"
"backup st.aj"
```

From the script it goes that the data are put into the Aladin stack that is stored in MySpace.

Phase 2. Result exploration by a researcher.

To make this phase technically is required:

- to run Aladin (3.030_votech release);
- to run Workbench;
- to open the Aladin stack stored in MySpace on the previous phase;
- to look through, analyze and perhaps edit the images obtained applying Aladin instruments.

An example on Fig. 5 shows what a researcher can see opening the Aladin stack. Here it is possible to see optical image with superposition of radio images obtained from FIRST (blue contours) and NVSS (red contours).

7 RVO Community Centre

One of the first steps of implementation of RVOII is organization of a Community centre in Moscow (at IPI RAS) for support of scientific astronomical problem solving over distributed repositories of astronomical information (containing data of observations, problem solving results, services for data and knowledge analysis). This Centre is positioned at the top layer of RVOII providing for its immediate usage for problem solving by scientists in astronomy.

The Centre has been created in October 2005 as an installation of the AstroGrid 1.1, developed recently in the UK and generously provided by the authors to be used for RVO. At present time the Centre includes two installations of Astrogrid 2006.3: an installation at the

IPI RAS and an installation at the Joint Supercomputer Centre (JSCC) of RAS [22]. The installation at the JSCC is aimed to provide more powerful and maintainable hardware basis for the Centre. The Centre is ready for use by astronomers including registering of Russian resources (data and services), establishing the required local data bases at the Centre, access to various information sources accumulated around the world, scientific problem solving applying workflow and program creation facilities of the AstroGrid. The instruction of how to start using it is provided at [3].

The Community Centre is considered as a part of the emerging International Virtual Observatory (Fig. 6).

Further plans of extending the Centre functionalities consist in the integration of subject mediator architecture [11] in the AstroGrid infrastructure, adding to it the data mining facilities, applying AstroGrid capabilities to implement different layers of RVOII.

References

[1] The Aladin Sky Atlas. http://aladin.u-strasbg.fr/

[2] AstroGrid Release 2006.2. http://software.astrogrid.org/

[3] AstroGrid of RVO as a Community Centre for astronomical problem solving over distributed information repositories accumulated in the world is ready to be used. (in Russian) http://synthesis.ipi.ac.ru//synthesis/projects/astromedia/astroannounce

[4] Briukhov D.O., Kalinichenko L.A., Zakharov V.N Diversity of domain descriptions in natural science: virtual observatory as a case study // Proceedings of the 7th Russian Conference on Digital Libraries RCDL'2005, Yaroslavl, Russia, 2005.

[5] Briukhov D.O., Kalinichenko L.A., Zakharov V.N., Panchuk V.E., Vitkovsky V.V., Zhelenkova O.P., Dluzhnevskaya O.B., Malkov O.Yu., Kovaleva D.A Information Infrastructure of the Russian Virtual Observatory (RVO). Second Edition, IPI RAN, 2005.

[6] L. Candela, D. Castelli, P. Pagano, and M. Simi. From Heterogeneous Information Spaces to Virtual Documents // In Proceedings of the 8th International Conference on Asian Digital Libraries,ICADL 2005, Bangkok, Thailand, December 2005, pages 11–22. Springer, 2005.

[7] Leonardo Candela, Donatella Castelli, Pasquale Pagano, and Manuele Simi. OpenDLibG: Extending OpenDLib by exploiting a gLite Grid Infrastructure // Proceedings of the 10th European Conference on Digital Libraries, ECDL2006, 2006 (in print).

[8] EGEE. Enabling Grids for E-science in Europe. http://public.eu-egee.org

[9] gLite. Ligthweight Middleware for Grid Computing. http://glite.web.cern.ch/

[10] M. Hoey. Lexical priming: a new theory of words and language. Routledge, London, 2005.

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

[12] 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.

[13] C. Lagoze, S. Payette, E. Shin, and C. Wilper. Fedora: An Architecture for Complex Objects and their Relationships // *Journal of Digital Libraries, Special Issue on Complex Objects*, 2005.

[14] R. R. Larson and R. Sanderson. Grid-based digital libraries: Cheshire3 and distributed retrieval. In Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries Digital Libraries: Cyberinfrastructure for Research and Education (JCDL05), New York, 2005. ACM.

[15] T. A. Phelps and P. B. Watry. A no-compromises architecture for digital document preservation // In Research and Advanced Technology for Digital Libraries 9th European Conference, ECDL2005, Proceedings, pages 266-277, 2005.

[16] A. Rajasekar,M.Wan, R.Moore,W. Schroeder, G. Kremenek, A. Jagatheesan, C. Cowart, B. Zhu, S.-Y. Chen, and R. Olschanowsky. Storage Resource Broker - Managing Distributed Data in a Grid // *Computer Society of India Journal, Special Issue on SAN*, 33(4):42–54, October 2003.

[17] 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.

[18] Russian Virtual Observatory. http://www.inasan.rssi.ru/rus/rvo/

[19] SDSS site. http://www.sdss.org/

[20] R. Tansley, M. Bass, and M. Smith. DSpace as an Open Archival Information System: Current Status and Future Directions // In *Proceedings of the 7th European Conference, ECDL 2003, Trondheim, Norway, August 2003*, pages 446–460. Springer-Verlag, 2003.

[21] IVOA Astronomical Data Query Language Version 1.01. http://www.ivoa.net/Documents/WD/ADQL/ADQL-20050624.pdf

[22] Joint Supercomputer Centre of RAS Web site, 2006. http://www.jscc.ru/

♣ This work was partially supported by the RFBR grants 05-07-90413-в and 06-07-89188-а.