APPLICATION DRIVEN MEDIATION MIDDLEWARE OF THE RUSSIAN VIRTUAL OBSERVATORY FOR SCIENTIFIC PROBLEM SOLVING OVER MULTIPLE HETEROGENEOUS DISTRIBUTED INFORMATION RESOURCES

D Briukhov[1], L Kalinichenko[1], D Martynov[1], N Skvortsov[1], S Stupnikov[1], A Vovchenko[1], V Zakharov[1], O Zhelenkova[2]

[1]Institute of Informatics Problems, Russian Academy of Sciences, Vavilov st. 44-2, Moscow, Russia
E-mail: brd@ipi.ac.ru, leonidk@synth.ipi.ac.ru, domartynov@gmail.com, nskv@ipi.ac.ru, ssa@ipi.ac.ru, itsnein@gmail.com
[2]Special Astrophysical Observatory, Russian Academy of Sciences, Nizhnij Arkhyz, Zelenchukskaya, Karachaevo-Cherkesia, Russia
E-mail: olbo@nm.ru

ABSTRACT

The paper[1] considers the middleware architecture of subject mediators in the hybrid grid-infrastructure of the Russian virtual observatory (RVO) for scientific problem solving over a set of heterogeneous distributed information resources (such as databases, services, ontologies) integrated by the mediators. The RVO hybrid infrastructure is constructed as a merge of the AstroGrid VO system developed in the UK and of the middleware supporting subject mediators developed at the Institute of Informatics Problems of RAS. An example of implementation in the hybrid architecture of a subject mediator for support of distant galaxy discovery problem is presented.

KEYWORDS: application driven architecture, mediation middleware, subject mediators, virtual observatory, problem solving, heterogeneous distributed information resources

1 INTRODUCTION

In various areas of science an exponential growth of the volume of experimental (observable) data obtained is revealed. Such data are obtained by a number of autonomous organizations, observable objects are diverse, technologies of observations are rapidly enhanced causing changes of structure and content of the collected information. These circumstances as well as rapid development of new software services for information processing during problem solving lead to enlarging of a gap between researchers and numerous data resources as well as software services. New approaches are required for information systems development focusing on specific facilities for problem solving over the set of distributed information resources accumulated in various scientific centers. New infrastructures including SOA, data grids, various interoperability providing middlewares are still far from solving semantic problems inherent to the integration of heterogeneous distributed resources (of data and services).
This paper is focused on the consideration of the astronomical Virtual Observatory (VO) infrastructures. The infrastructure of the Russian VO (RVO) (Briukhov & Kalinichenko & Zakharov & Malkov & Kovaleva & Zhelenkova, 2005) is based on the IVOA standards (Williams & Hanisch & Linde, 2004) enhanced with the subject mediation concept providing for application domain conceptual definition for formulation and solving of various classes of scientific problems in terms of the concepts of such domains, information entity structures, services and processes defined in a declarative way. Mediators provide key role for solving of semantic problems of heterogeneous resources integration. In particular, during the resource integration a mediator should semantically match specifications represented in various information models and provide semantically correct mapping of schemas of integrated resources into the mediator schema. Due to heterogeneity of resources represented in different information models for the uniform representation of their semantics during resource integration it is required to bring various information models to the unified representation in frame of a common information model called the *canonical* one. The canonical information model serves as a common language (“Esperanto”) for adequate representation of semantics of heterogeneous information representation models used in information resources relevant to the mediator. The methods for information models mapping and for synthesis of canonical information models for subject mediator middleware are considered in (Kalinichenko & Stupnikov, 2008).

[1] This research has been done under the support of the RFBR (projects 06-07-89188-a, 08-07-00157-a) and the Program of the Department of Nanotechnologies and Information Technologies of RAS entitled as Basic principles of information technologies and systems (project 1-10).

Another set of semantic problems of development of mediators for VO includes problems related to the subject mediator specification and registration of relevant resources in it. Specification of a subject mediator for a class of problems includes definitions of application domain concepts expressed by the respective ontologies, definition of application domain classes, definition of instance types of such classes and their methods, definition of problem solving processes specifying concurrent activities implemented by class methods, services and other processes. Such specifications are expressed in the canonical model having formal semantics. The result of the subject mediator specification provided by a respective scientific community as a consensus reached. The activity for the mediator specification is called the period of its *consolidation*.
Registration of relevant to the mediator resources is considered similarly to the compositional system development. Resource registration in the mediator is a process of purposeful transformation of specifications including decomposition of the mediator specification into consistent fragments, search among specifications of relevant resources of the adequate data types treated as candidates for refining by them of the mediator data type specifications, construction of expressions defining resource classes as a composition of the mediator classes.
For this paper it is important that in the UK the AstroGrid system is being developed (AstroGrid, 2008) as a complete infrastructure for support problem solving in VO. AstroGrid services in general conform to the functions required for implementation the RVO information infrastructure (RVOII). However, AstroGrid does not provide advanced facilities for problem solving over sets of heterogeneous information resources such as subject mediators. Due to that in 2006 under support of RFBR (the Russian Foundation for Basic Research) a project of the hybrid architecture development merging capabilities of AstroGrid with the architecture of subject mediators was initiated. The objective of this paper is a brief survey of results of development of such hybrid architecture focusing on new methods and facilities for specification of application domains and information resources and problem solving in the hybrid infrastructure. As far as the development of the hybrid infrastructure is mainly related to the interfacing of run time facilities of both infrastructures, the paper is focused mostly on them.

2 PROBLEM SOLVING APPROACH OVER HETEROGENEOUS DISTRIBUTED INFORMATION RESOURCES

The approach to integrated representation of subject area of a problem considered in this paper is *application-driven* with respect to a set of information resources relevant to the problem. This means that a description of an application subject area is created independently of the resources and after that the resources relevant to the application are mapped into this description. The approach assumes creation of a subject mediator that supports an interaction between the application and resources on the basis of application domain definition (i.e. description of the mediator).
To reach heterogeneous Grid-based information resources convergence the mediators provide a common framework through which to resolve the mediation challenges:

- canonical information model construction for unified definition of heterogeneous ontologies, data, services;
- mediator consolidation;
- relevant heterogeneous resources identification and semantic integration in mediator;
- semantic support of the canonical and mediator models as well as information resource models;
- application problems specification and interpretation.

Specific mediation middleware developed includes

- extensible canonical information model to specify mediators and heterogeneous information resources;
- heterogeneous information models Unifier for canonical information models construction;
- facilities for relevant resources discovery and their semantic registration in mediators;
- facilities for application problem specification;
- facilities interpreting problem specifications in the mediator context over the registered resources.

Further the constituent parts of the mediation middleware are described briefly.
The main principle of *canonical model synthesis* consists in its extensibility required to reach semantic integration and information interoperability in environments that include various heterogeneous models. A kernel of the canonical model is fixed. For each specific information model *M* of the environment an extension of the kernel is defined so that this extension together with the kernel is refined by *M*. Such refining transformation of models should be provably correct. To formalize the notion of *refinement* the Abstract Machine Notation and B-Technology (Abrial, 1996) is applied. The canonical model for the environment is synthesized as the union of extensions, constructed for models *M* of the environment. The canonical information model synthesis method that we have developed firstly for the structured data models provided a seminal role for synthesis of canonical models for various kinds of resource information models including object models, service models, ontological and process models (Kalinichenko, 2004).
A prototype of the *Heterogeneous Information Model Unifier* aimed at partial automation of methods of the canonical models synthesis has been constructed (Kalinichenko et al., 2008). One of the practical purposes of this work is to support the heterogeneous information resource integration in a specific subject mediator. Due to that, the

subject mediator specification information model (the SYNTHESIS language (Kalinichenko & Stupnikov & Martynov, 2007) - hybrid semistructured and object oriented information model) has been chosen as the canonical model kernel. The aim of the Model Unifier is to unify a set of information models used interoperably in some IS. A resource model *R* is said to be *unified* if it is mapped into the canonical model *C*. This means a creation of such extension *E* of the canonical model kernel and such mapping *M* of a resource model into extended canonical one that the resource model *refines* the extended canonical one. Model refinement of *C* by *R* means that for any admissible specification *r* represented in *R* its image *M(r)* in *C* under the mapping *M* is refined by the specification *r*. Process of model mapping includes a possibility of proving that arbitrary specification *r* represented in *R* refines its image *M(r)*. Verification of model refinement is realized over a set of resource model specification samples. Model Unifier makes process of information model mappings (that is quite hard) semi-automatic. Special third-party facilities (Meta-Environment (Van den Brand & van Deursen & Heering, 2001)) are used in Unifier for declarative specification of such mappings (compilers) and generating them according to the meta specifications.
The Unifier is considered as a constituent part of the subject mediator middleware.
A process of registration of heterogeneous information resources in a subject mediator is based on GLAV technique (Briukhov & Kalinichenko & Martynov, 2007) that combines two approaches - LAV (Local As View) and GAV (Global As View). According to LAV the schemas of resources being registered are considered as materialized views over virtual classes of the mediator. GAV views provide for reconciliation of various conflicts between resource and mediator specifications and provide rules for transformation of a query results from resource into mediator representation. Such registration technique provides for stability of application specification during any modifications of specific information resources and of their actual presence (removing, addition new ones, etc.) as well as for scalability of mediators w.r.t. the number of resources registered in them.
A principal point of this approach is formal proof of the refinement of mediator specification fragments by resources specifications during the process of construction of specification mappings. The techniques listed are used as a basis for the tool prototype for identification and registration of information resources in mediator.
Mediator middleware (Figure 1) contains also facilities for problem specification and interpretation in the mediator context over the registered resources. To save space, the details of Model Unifier and Resource Registration Tool are omitted.

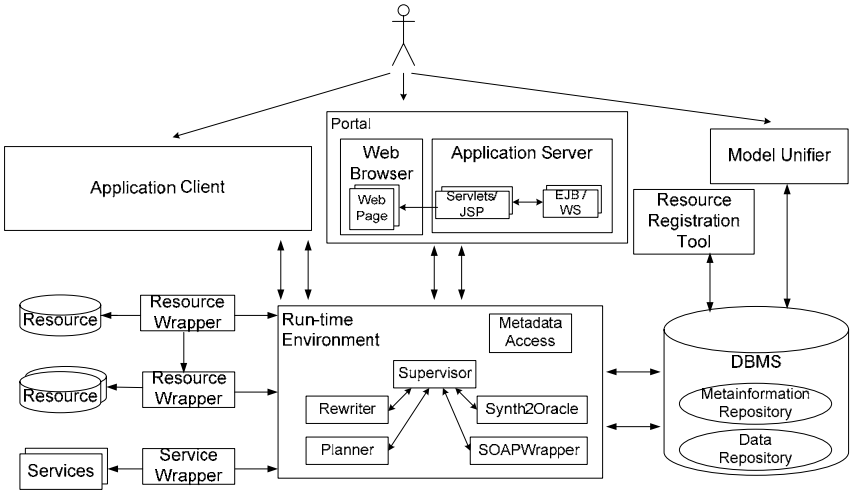

Figure 1. Mediator Middleware Architecture

Application Client provides for a graphical user interface of the mediator supporting mediator canonical program (query) language and convenient representation of query results. *Portal* is an Internet equivalent of the Application Client.
Run-time environment provides for user program (query) processing; program rewriting into queries over registered resources, transmitting rewritten queries to respective wrappers (adapters); receiving results from adapters; transmitting query results to a user.
Wrappers connect a mediator to the concrete resources. Wrappers transform queries formulated in mediator language into languages of concrete resources and conversely transform results of queies represented in formats of resources into the format of mediator.
Metainformation and Data repository is used as a storage for mediator metainformation (used for query transformation) as well as a storage for intermediate results of queries returned by resources.

3 ASTROGRID AS A BASIS OF RVO INFRASTRUCTURE

The AstroGrid system is aimed at support of infrastructure for scientific problem solving in astronomic virtual observatories. The AstroGrid provides access facilities to astronomic catalogs, digital surveys and image archives as well as to metadata registries (in which VO resources are registered). The AstroGrid system is based on Web-service architecture. Main components of the AstroGrid are the following:

- *Registry* is a collection of metadata that are represented as XML-documents describing resources which can be used during problem solving by means of VO. Registry is implemented on the base of the OAI PMH standard, specialized by IVOA for VOs. Applications, services, information resources, AstroGrid components may be used as resources. Registry supports *harvesting* of resource metadata, registered in another registries of VO.
- *VOSpace* is a virtual data storage. Any AstroGrid service can access VOSpace. The component is used for storing and transmitting files among services solving one or multiple problems in frame of VO.
- *Common Execution Architecture (CEA)* component defines a way of converting an application into AstroGrid service. CEA-application can be called from workflows. Also CEA-application can be called as stand-alone application. It can be registered and discovered applying AstroGrid registries, it can access VOSpace, etc.
- *Data Set Access (DSA)* component implements a connection of a database to AstroGrid. DSA provides two access interfaces: CEA-application executing queries to the database and Cone Search allowing to search data in databases according to some specific area of the sky given as a parameter.

The analysis shows that applying the AstroGrid as a basis of the RVO infrastructure allows to implement the principles of RVOII construction, namely: support of grid-interoperable services, modularity of the architecture, possibility of reuse and composition of services, creation of multilayered architecture. AstroGrid components are directly applicable as a core of RVOII architecture.

4 HYBRID GRID INFRASTRUCTURE FOR PROBLEM SOLVING

The main goal of the hybrid grid infrastructure is to provide interoperability of the mediator middleware and AstroGrid System. Hybrid infrastructure (shown on Figure 2) provides the following important capabilities:

- registration and search of subject mediators in AstroGrid registries;
- mediator schema and other metadata lookup by users;
- querying mediators applying the mediator query language (Syfs) or ADQL (Ohishi & Szalay, 2005);
- using mediator functions as steps of AstroGrid workflows;
- browsing of the result returned by mediator by means of native AstroGrid interface.

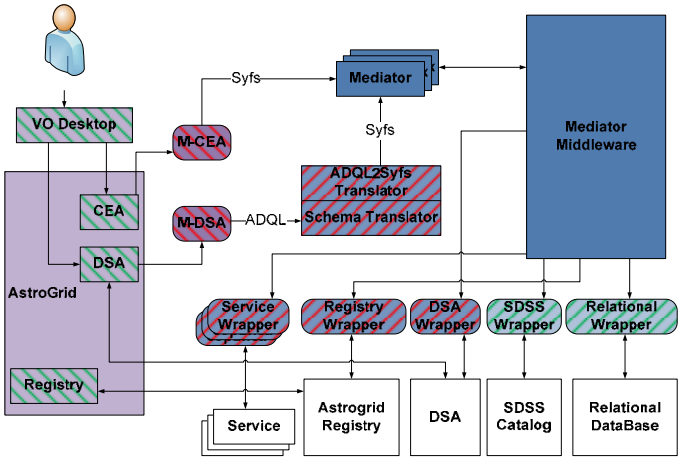

Figure 2. Hybrid architecture of AstroGrid System and Mediator Middleware

Mediators can be registered in AstroGrid registries as CEA-applications as well as DSA components. Mediator registered as CEA-application is used to submit queries expressed in the mediator query language in terms of mediator schema. Mediator registered as DSA is used to submit queries expressed in ADQL query language in terms of relational view over a mediator schema. In both ways a result is returned in format compatible with AstroGrid. To provide this capability *M-CEA* and *M-DSA* components were developed encapsulating mediator's instance. *Schema Translator* and *ADQL2Syfs Translator* components were developed to provide an opportunity to write queries in relational query language ADQL in terms of mediator schema (which is object-oriented). *Schema Translator* is used to create a relational view of a mediator schema. *ADQL2Syfs Translator* is used to translate ADQL query in terms of

relational view of mediator schema into Syfs query in terms of mediator schema. Both services are used in *M-DSA* component. Representation of a mediator as DSA component is very important advantage, because AstroGrid user may use mediator in a native way, just as ordinary DSA resource.
Wrappers are used to translate queries expressed in the mediator query language into queries in a resource query language, to submit query to resources and to receive results and return them in the AstroGrid-compatible format.
Two specific wrappers have been developed: the wrapper of DSA resources connecting them to the mediator, and the wrapper of AstroGrid registries (to realize metadata search by the mediator). Also Relational Wrapper and SDSS Wrapper with XMatch capability have been developed allowing to use any kind of relational database, SDSS catalog, any DSA catalog or registry as resources in hybrid grid infrastructure prototype.

5 DISTANT GALAXY SEARCH PROBLEM

As an example of hybrid architecture application, the distant galaxy search problem has been experienced recently.
This astronomic problem is a distant galaxy search in the sky strip investigated in the "Cold" deep survey with the RATAN-600 (large Russian radio telescope). *RC* catalogue (selected from the survey (Parijskij & Goss & Kopylov & Soboleva & Temirova & Verkhodanov et al., 1996)) was used as a list of initial radio sources. Optical sources were selected from DR 3 SDSS catalogue and cross-matched with the RC catalogue. The result of the cross-match may contain candidates for distant galaxies that should be analyzed further applying their images and special astronomic tool Aladin capabilities.
Use of astronomical resources was a weak point of the previous AstroGrid application. Firstly, addition of new resources was quite hard in that environment. Secondly, *XMatch* (cross-match implementation) is a remote application, so it is required to transfer data through the network, and in case of SDSS the required amount of data is huge. These problems were solved with the help of mediators in hybrid infrastructure by replacing steps in which distant galaxy candidates were searched. Extraction of images has been done also by means of Aladin. Informal workflow for Distant Galaxy Search problem looks as follows:

- Extract data from RC catalog in some area;
- Extract data from SDSS catalog in same area as in RC catalog case;
- Cross-match results of previous steps;
- Chose the optical sources satisfying following rules: (u+r)/2 – g > 1; (g+i)/2 – r > 1; (r+z)/2 – i > 1 (these are restrictions on spectral indices u, r, g, i, z taken from *PhotoPrimary* SDSS class specification);
- Display image to user for defined radio and optical sources.

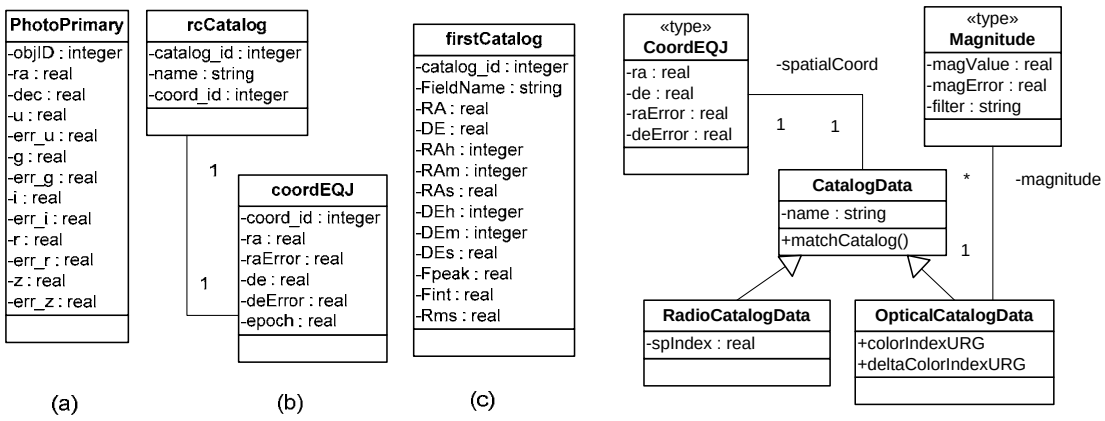

Figure 3. Resource schema and mediator schema specifications (fragments)

Mediator and resources schema specifications fragments for the distant galaxy search problem are shown on Figure 3. The respective mediator has been created and resources have been registered in it. Mediator was also registered as CEA-application with Syfs query interface. For distant galaxy search problem as a whole a workflow was created in the AstroGrid system. This workflow consists of two main parts: distant galaxy candidates search by means of mediator, and image retrieval by means of Aladin. Query to mediator intended to find distant galaxy candidates looks as follows:

```
{{ r(x/[ra, de, name, name1, ra1, de1])
:-radioCatalogData(y/[name, ra: spatialCoord.ra, de: spatialCoord.de])
& opticalCatalogData(x/[name1: name, ra1: spatialCoord.ra, de1: spatialCoord.de, colorIndexURG])
& matchCatalog(y, x, 45, 45, b) & b = true
& ra >= 120.0 & ra <= 255.0 & de >= 4.39 & de <= 5.61
& ra1 >= 120.0 & ra1 <= 255.0 & de1 >= 4.39 & de1 <= 5.61
```

& colorIndexURG > 1 }}
Query consists of one rule which is the main part of distant galaxy search problem solution. Result of this query consists of astronomical sources, discovered in radio and optical catalogs, cross-matched and satisfied the constraints for color indices. Result is returned in AstroGrid-compatible format (a table each row of which is a distant galaxy candidate).
Aladin image retrieval tool is used to retrieve images for distant galaxy candidates from optical image archives. Aladin is a sky atlas providing visualization of digital astronomical images, superposition of data from astronomical catalogs and databases, access to astronomical databases. Aladin provides graphical user interface as well as program interface throw Aladin script language. To retrieve all images in AstroGrid a CEA-application was developed, which provides a possibility to execute programs written in the script language. Result is presented as Aladin stack and can be opened from Aladin. This application is invoked for each distant galaxy candidate in a loop. The aim of an expert is to decide whether the candidate combination of images contains distant galaxy or not.

6 CONCLUSION

In this paper the first results of development of the subject mediator middleware in the hybrid grid-infrastructure of VO for problem solving over the set of heterogeneous distributed information resources are presented. Such middleware construction is oriented on resolving of a number of semantic challenges imposed on a scientist who should solve a problem in some application domain over various relevant results of observations and facilities for their processing accumulated in the world. The hybrid VO architecture is implemented as an infrastructure integrating of the AstroGrid VO developed in the UK and facilities for support of the subject mediator middleware created at the IPI RAS. In the subject mediator architecture an approach driven by the applications is implemented according to which for the class of problems of an application a specification of the application domain is specified independently of the existing information resources. An example of simple subject mediator for support of distant galaxy discovery in the integrated architecture is given. The results obtained evidence that the approach chosen is prospective. Further extension of the approach is planned for solving of various problems at RVO.

7 REFERENCES

Abrial, J.-R. (1996) The B-Book: Assigning Programs to Meanings. Cambridge, UK: Cambridge University Press.

AstroGrid Project (2008) Retrieved October 24, 2008 from the WWW: http://www.astrogrid.org

Briukhov, D.O. & Kalinichenko, L.A. & Zakharov, V.N. & Zhelenkova, O.P. & Malkov, O.Yu. & Kovaleva, D.A. (2005) Information Infrastructure of the Russian Virtual Observatory (RVO). Moscow, Russia: IPI RAN.

Briukhov, D. O. & Kalinichenko, L. A. & Martynov, D. O. (2007) Source Registration and Query Rewriting Applying LAV/GAV Techniques in a Typed Subject Mediator. Proc. of the Ninth Russian Conference on Digital Libraries RCDL'2007 (pp. 253-262). Pereslavl-Zalesskij, Russia.

Kalinichenko, L.A. (2004) Canonical model development techniques aimed at semantic interoperability in the heterogeneous world of information modeling. Proceedings of the Open INTEROP Workshop “Enterprise modeling and ontologies for interoperability” at the 16th CAiSE Conference (pp. 101–116). Riga, Latvia.

Kalinichenko, L.A. & Stupnikov S.A. (2008) Constructing of Mappings of Heterogeneous Information Models into the Canonical Models of Integrated Information Systems. Advances in Databases and Information Systems: Proc. of the 9th East European Conference (pp. 106-122). Pori, Finland.

Kalinichenko, L. A. & Stupnikov, S. A. & Martynov, D. O. (2007) SYNTHESIS: a Language for Canonical Information Modeling and Mediator Definition for Problem Solving in Heterogeneous Information Resource Environments, Moscow, Russia: IPI RAS.

Ohishi, M. & Szalay A. (2005) IVOA Astronomical Data Query Language. Retrieved October 24, 2008 from the WWW: http://www.ivoa.net/Documents/WD/ADQL/ADQL-20050624.pdf

Parijskij, Yu. & Goss, W. & Kopylov, A. & Soboleva, N. & Temirova, A. & Verkhodanov, O. & Zhelenkova, O. & Naugolnaya, M. (1996) Investigation of RATAN-600 RC radio sources. Bulletin of SAO, 40, 5-124.

Van den Brand, M. G. J. & van Deursen, A. & Heering, J. (2001) The ASF+SDF meta-environment: a component-based language development environment. In Wilhelm, R. (Ed.) Compiler Construction 2001, Berlin, Germany: Springer-Verlag.

Williams, R. & Hanisch, B. & Linde, T.(2004) Virtual Observatory Architecture Overview. Retrieved October 24, 2008 from the WWW: http://www.ivoa.net/Documents/Notes/IVOArch/IVOArch-20040615.html