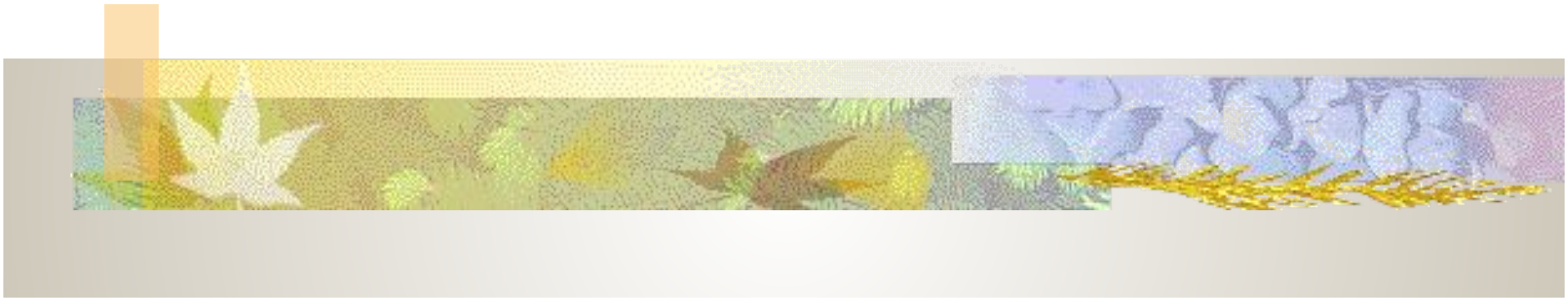
Интеграция информации для решения задач над множеством неоднородных распределенных информационных ресурсов

Спецсеминар факультетов ВМК и Мехмата МГУ для студентов 1 – 5 курсов

Вводное заседание:

15 октября, 18:30; ауд. 371 ВМК

Вместо введения



Семинар ведут:

- Проф. ВМК, зав. лаб. Института проблем информатики РАН, д.ф.-м.н. Леонид Андреевич **Калиниченко** (e-mail: leonidk@synth.ipi.ac.ru)
- Проф. ВМК, зав. лаб. Открытых информационных технологий ВМК МГУ, д.т.н. Владимир Александрович **Сухомлин** (e-mail: sukhomlin@mail.ru)
- С.н.с. лаб. Логических проблем информатики Мехмата МГУ, с.н.с. Математического института им. Стеклова РАН, к.ф.-м.н. Ростислав Эдуардович **Яворский** (e-mail: rey@mi.ras.ru)
- С.н.с. Института проблем информатики РАН, к.т.н. Сергей Александрович **Ступников** (e-mail: ssa@ipi.ac.ru)

Содержание семинара

- Изучение активно развиваемых перспективных **технологий создания распределенных интегрированных информационных систем** в **совокупности с необходимыми в таких технологиях формальными методами**
- Идея семинара: ориентация студентов на эту область как **возможную профессию**, введение студентов в состояние соответствующей проблематики в мире
- Главное – достижение студентами такого уровня, чтобы они могли **практически участвовать в реальных проектах** (конкретные задания и условия участия в проектах: в рамках курсовых и дипломных работ, статей, разработок, диссертаций, и др.)
- Нет стремления ограничивать семинар участием студентов определенных курсов – **важен интерес к проблематике и проявляемая инициатива**. Вписывание этой работы в учебный план – вопрос, который подлежит уточнению.

Содержание семинара (2)

- Проблемы создания новых **методов и инфраструктур для решения задач над множеством неоднородных распределенных информационных ресурсов**, проблемы **формального определения разнообразных предметных областей**, проблемы **семантической интеграции информационных ресурсов**, проблемы **синтеза канонических информационных моделей для такой интеграции**, и пр. Рассмотрение этих вопросов будет сопровождаться **изучением существующих и разрабатываемых формальных методов** и подходов для разрешения названных проблем.
- Семинар основан на **значительном опыте применения формальных методов при создании новых подходов к решению задач над множеством неоднородных распределенных ресурсов**, накопленном в Лаборатории Методов композиционного проектирования информационных систем ИПИ РАН и в Лаборатории Открытых информационных технологий ВМК МГУ. **С некоторыми проектами и публикациями можно познакомиться на сайте http://synthesis.ipi.ac.ru/synthesis**

Свежая информация (от 23.09)

- **Communications of the ACM**,

  Volume 51, Number 9 (2008), Pages 54-59
- **Software engineering and formal methods**

  Mike Hinchey, Michael Jackson, Patrick Cousot, Byron Cook, Jonathan P. Bowen, Tiziana Margaria

- **Communications of the ACM.**

  Volume 51, Number 9 (2008), Pages 72-79
- **Information integration in the enterprise**

  Philip A. Bernstein, Laura M. Haas

Множество распределенных неоднородных ресурсов в Интернете

- **Экспоненциальный рост** в Интернете числа неоднородных информационных ресурсов (баз данных, сервисов, процессов), разработанных для решения разнообразных задач
- **Неоднородность**: разномодельность, различная семантика, контекст …
- Потребность **совместного использования (интеграции)** модельно неоднородных информационных ресурсов
- Примеры неоднородных информационных моделей: **моделей данных** ODMG 2000, SQL 2006, UML, стеки XML и RDF моделей, **моделей потоков работ** Staffware, COSA, InConcert, Eastman, FLOWer, Domino, Meteor, Mobile, MQSeries, Forte, Verve, Vis.WF,Changeng, IFlow, SAP/R3, **языков процессной композиции сервисов** XPDL, BPEL, BPML, XLANG, WSFL, WSCI, семантических моделей (включая **онтологические модели**, **модели представления знаний**, **модели метаданных** и многие другие)
- Инфраструктуры, технически способствующие организации решения задач в таких условиях. Среди них **Веб-сервисы, Гриды данных, Семантический Веб, инфраструктуры интеграции информационных ресурсов, интероперабельные инфраструктуры промежуточного слоя, и др.**

Проблемы семантики при решении задач (примеры)

1. **Абстрактное определение предметных областей** и их понятий;
2. **Отображение и интеграция контекстов** предметных областей информационных ресурсов в контекст предметной области задачи;
3. **Идентификация релевантных задаче информационных ресурсов** и формировании их композиций;
4. **Доказательно правильное отображение информационных моделей ресурсов в информационную модель предметной области**;
5. **Интеграция схем ресурсов в схеме предметной области** и устранение разнообразных конфликтов;
6. **Выявление семантически подобных компонентов ресурсов** в процессе интеграции схем;
7. **Адекватное преобразование формул (запросов) программы решения задачи, выраженных в терминах предметной области задачи**, в формулы, выраженные в схемах релевантных ресурсов, и пр.

Для разрешения такого рода проблем и предназначены опирающиеся на **математическую логику и алгебру формальные методы**, представляющие собой совокупность языков, технологий и инструментальных средств спецификации и верификации интероперабельных систем (ИС). **Формальные методы достигли индустриальной зрелости.**

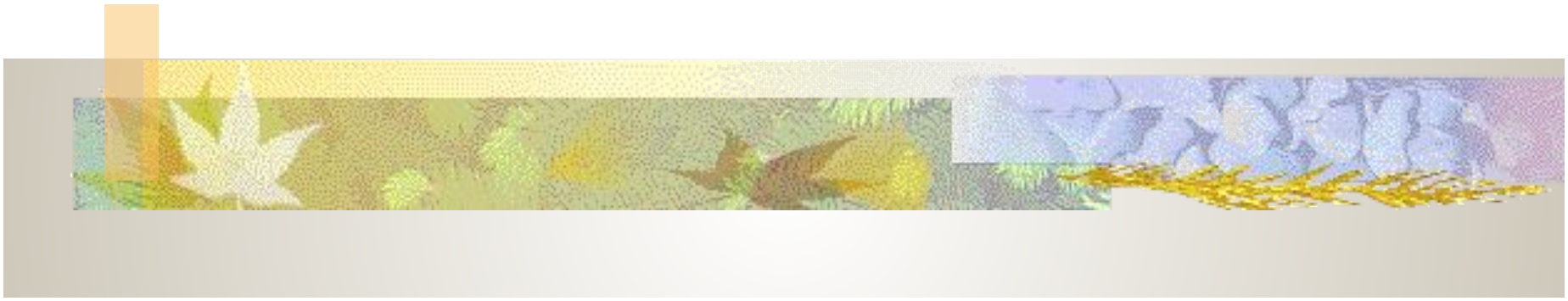
Примеры тем для изучения на семинаре

- Формальные **методы и языки спецификации и проектирования** ИС (такие как Нотация Абстрактных Машин, UML в формальном представлении, Объектный Z, и др.). Примеры их применения.
- Языки и **средства онтологического моделирования** предметных областей
- **Концептуальное моделирование** предметных областей
- Архитектуры средств **семантической интеграции неоднородных информационных ресурсов** для решения задач. Понятие и архитектуры **предметных посредников**.
- Методы **синтеза канонических информационных моделей**, позволяющих унифицировать множество информационных моделей (языков) представления информационных ресурсов. Построение отображений разнообразных моделей информационных ресурсов в каноническую с доказательством правильности отображения. Язык СИНТЕЗ.
- **Формальная семантика канонических моделей.**

Примеры тем для изучения на семинаре (2)

- **Методы идентификации информационных ресурсов**, максимальные фрагменты спецификаций которых доказательно уточняли бы фрагменты концептуального описания предметной области. Методы основаны на алгоритмах сопоставления (matching) спецификаций и **доказательстве достижения факта уточнения**
- **Методы интеграции онтологических (понятийных) контекстов предметной области и информационных ресурсов**. Методы основаны на доказательстве поглощения (subsumption) одного понятия другим, используя **дескриптивные логики**
- **Методы приведения процессных моделей** (широко использующихся для управления корпоративной деятельностью) **к унифицированной модели** (одной из практических целей является создание **виртуальных организаций**, объединяющих деятельность нескольких реальных организаций)
- **Методы построения композиций фрагментов спецификаций разнообразных ресурсов**, которые уточняли бы заданные спецификации предметной области. Эти задачи требуют проведения доказательств на основе логики предикатов.
- **Методы преобразования программ решения задачи**, выраженных в терминах предметной области, в программы над информационны-ми ресурсами с доказательством свойства включения полученного результата вычисления таких программ в ожидаемый результат.

Виртуальные обсерватории (ВО): пример решения задач в архитектуре предметных посредников



PROBLEM SOLVING APPROACH OVER HETEROGENEOUS DISTRIBUTED INFORMATION RESOURCES

- Mediation based, application driven approach, mediation middleware
- Mediator canonical model synthesis and heterogeneous information model unifier
- Registration of relevant resources in a mediator
- Mediation runtime environment

MEDIATION-BASED, APPLICATION DRIVEN APPROACH

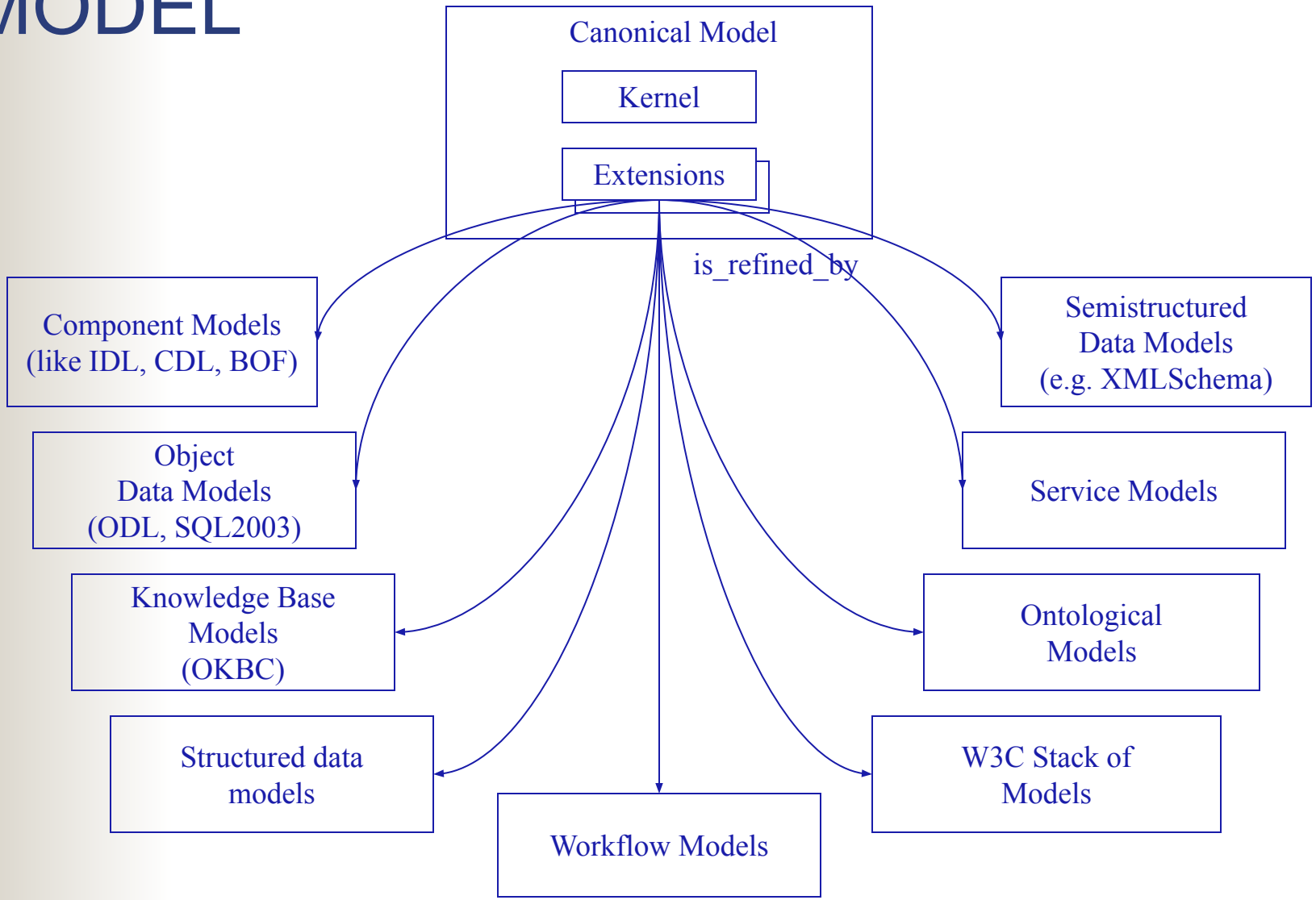
- ***Application-driven subject domain specification approach is emphasized***: Abstract **specification of subject domain is provided** in terms of concepts, data structures, functions, processes **independently of existing resources specifications**
- Domain specification consolidated by the respective community **is treated as a mediator**
- **Registration of relevant resources in mediator is based on their semantic mapping into the mediator**
- **To reach heterogeneous [Grid-based] information resources convergence** mediators provide **a common framework** from which to operate
- Mediation challenges: 1) **canonical information model** construction for unified definition of heterogeneous ontologies, data, services, processes; 2) **mediator consolidation**; 3) relevant heterogeneous **resources identification and semantic integration** in mediator; 4) **semantic support** of the canonical and mediator models, information resource models; 5) **application problems** specification and **interpretation**

MEDIATION MIDDLEWARE

The mediation middleware includes:

1. extensible **canonical information model** to specify mediators,
2. heterogeneous **information model Unifier** for canonical information models construction,
3. facilities for **relevant resources discovery** and their **semantic registration** in mediators,
4. facilities for **application domain specifications**,
5. facilities **interpreting problem specifications in the mediator context over the registered resources**.

HETEROGENEOUS MODELS ABSORBED BY THE CANONICAL MODEL



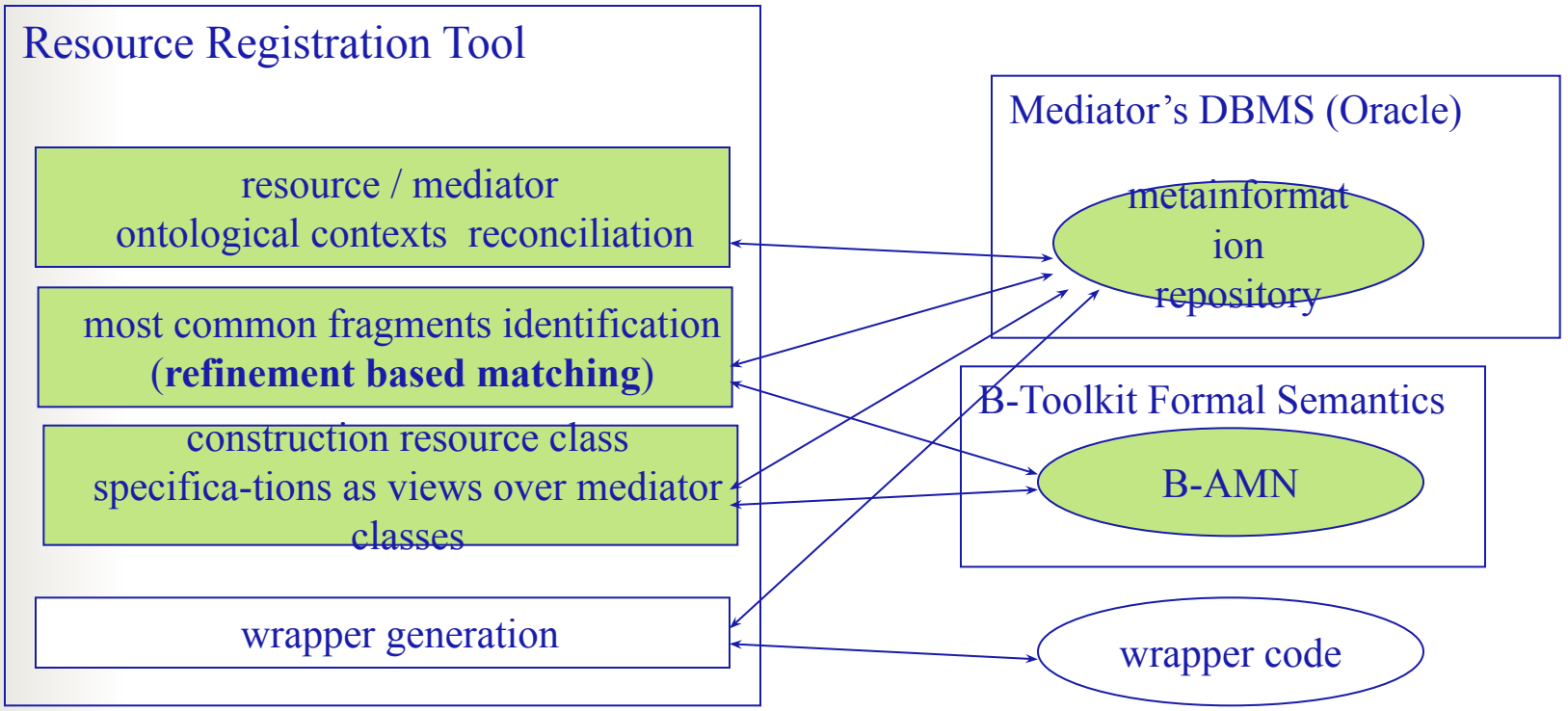
HETEROGENEOUS INFORMATION MODEL UNIFIER

- Hybrid semi-structured – object language is used as a **canonical model kernel**
- 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 **has been experienced for various kinds** of resource information models including **data models, service models, ontological models, process models**
- **Process of information model mappings (quite hard)** is now semi-automatic due to development of the **Heterogeneous Information Model Unifier** prototype that assists in construction of **mapping of a specific information model into the canonical one**
- We apply *Meta Environment* facilities providing for **declarative specification of such mappings** (compilers) and generating them according to the meta specifications

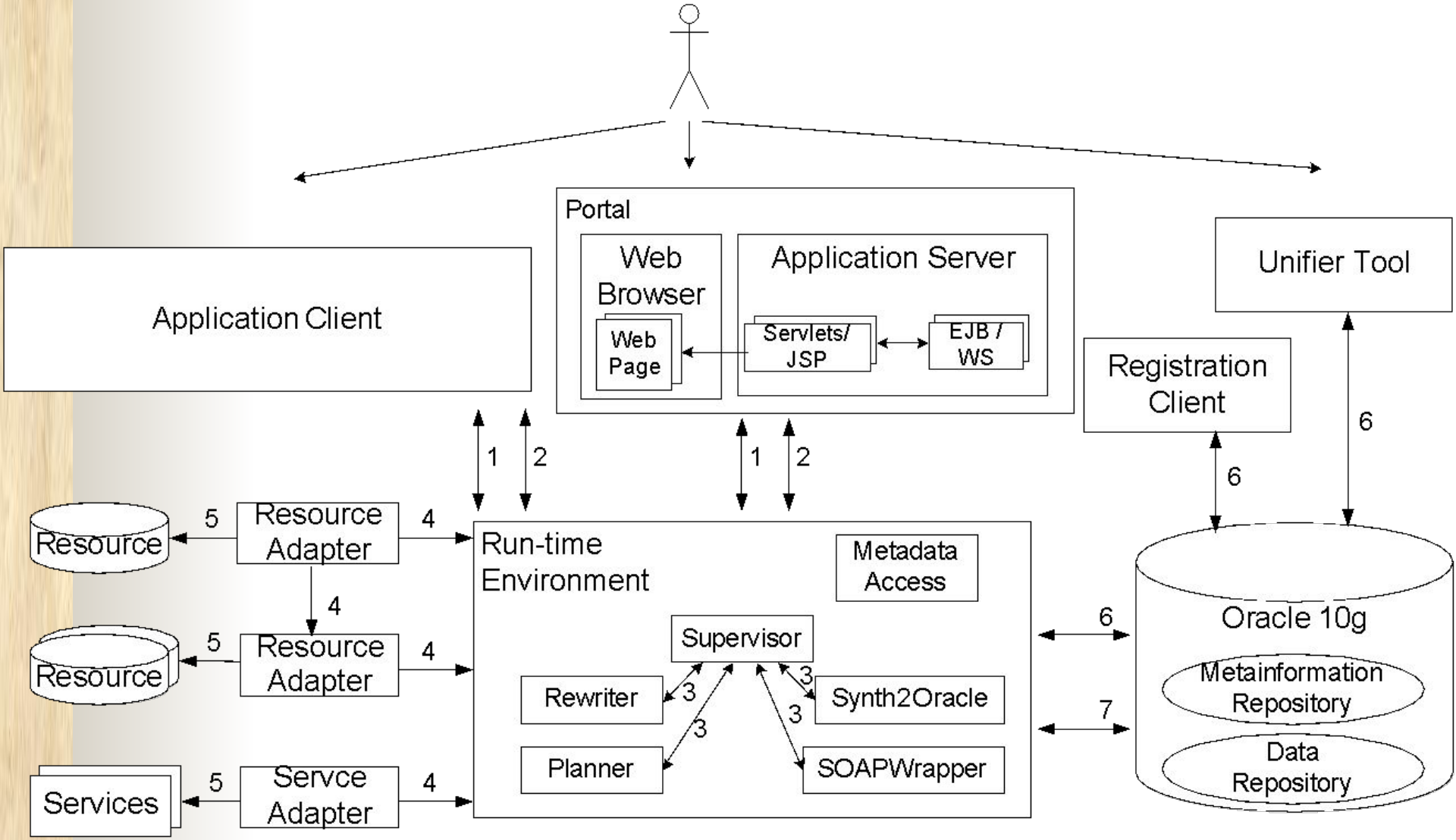
REGISTRATION OF RELEVANT RESOURCES IN THE MEDIATOR

- Grid-based resource specifications are uniformly represented in the canonical model applying mapping facilities generated by the Unifier
- Registration of resources in a mediator **is a process of** decomposition of **mediator specifications** into consistent fragments and discovery of refinement-based matching of resource and mediator specification fragments
- The main registration result is a *set of* ***GLAV – like*** *expressions* **defining how a resource class is determined as a composition of the mediator classes and functions**
- Identification and matching of relevant resource and mediator fragments are based on three models: **metadata model, ontological model, canonical information model**
- These techniques are used as a basis for the tool for registration of information sources in mediator

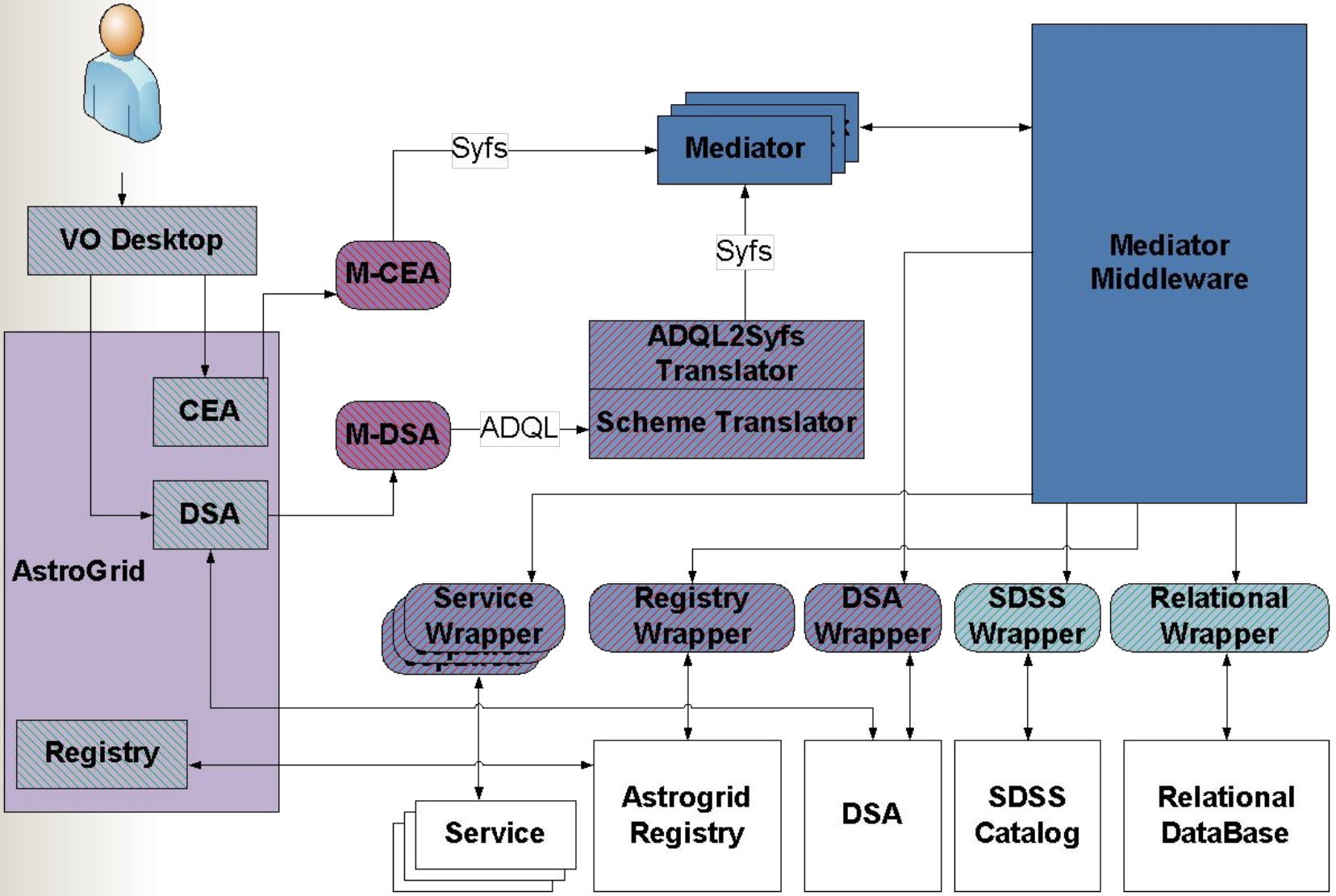
STRUCTURE OF THE RESOURCE REGISTRATION TOOL



MEDIATION MIDDLEWARE ARCHITECTURE



HYBRID INFRASTRUCTURE OF MEDIATORS AND ASTROGRID

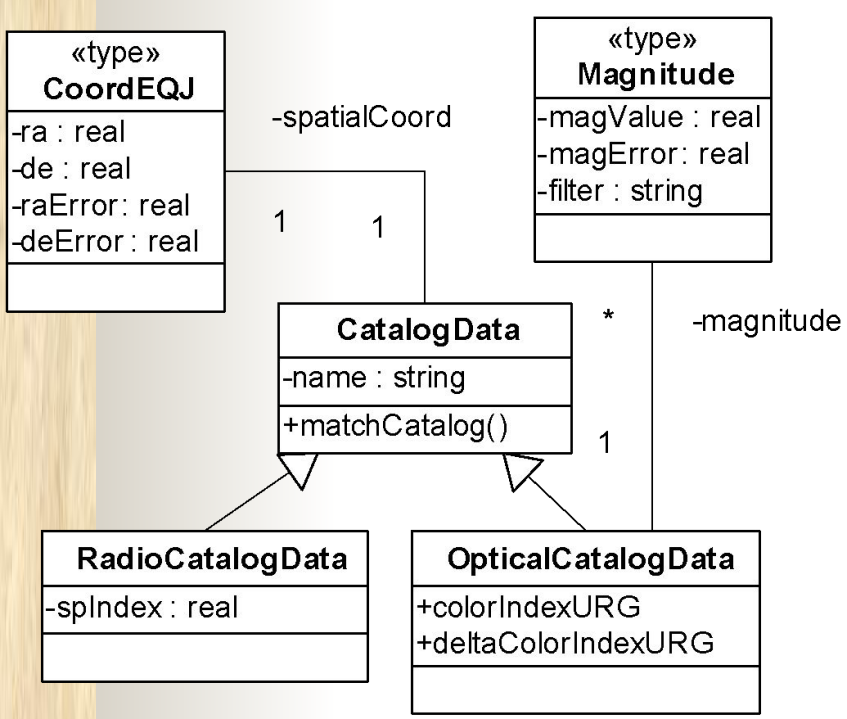


INFORMATION PROCESSING STEPS AT DISTANT GALAXY SEARCH

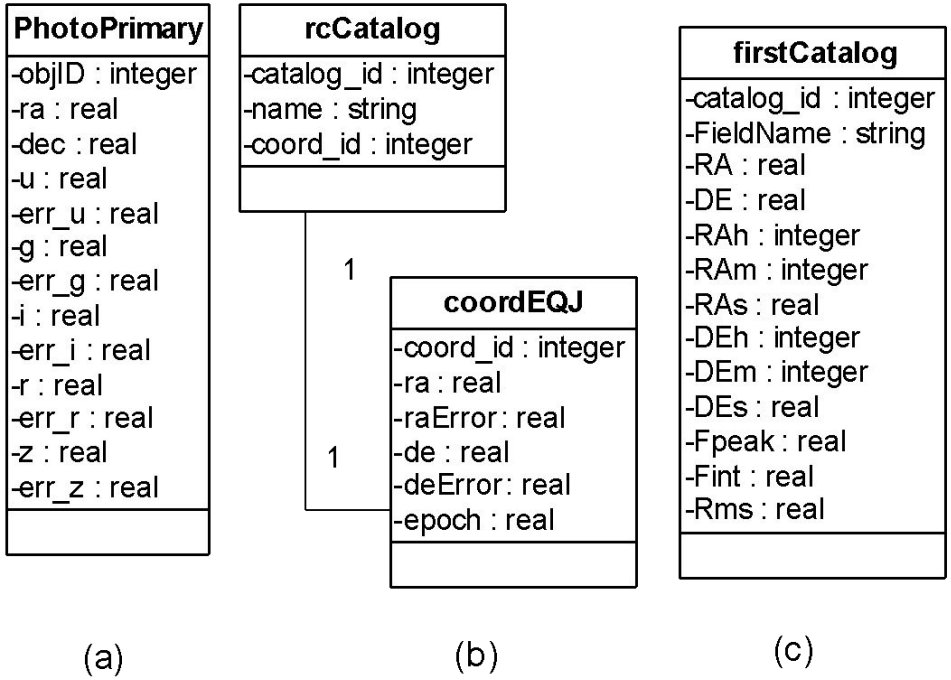
The first astronomical problem that has been supported by **IPI RAS** together with the **Special Astrophysical Observatory of RAS (SAO RAS)** applying mediation middleware, **AstroGrid** and **Aladin** is a distant galaxy discovery in the sky strip investigated in the "Cold" deep survey with the **RATAN-600**

1. Mediator part - search of distant galaxy candidates
2. Prepare data for Image Retrieval tool (workflow script)
3. Get images in a loop (GetImage), download the respective images from FIRST and SDSS images archives, and provide their superposition.
4. Analyze the result

STEP1: MEDIATOR SCHEMA (SIMPLIFIED)



Mediator Schema



Schemas of resources

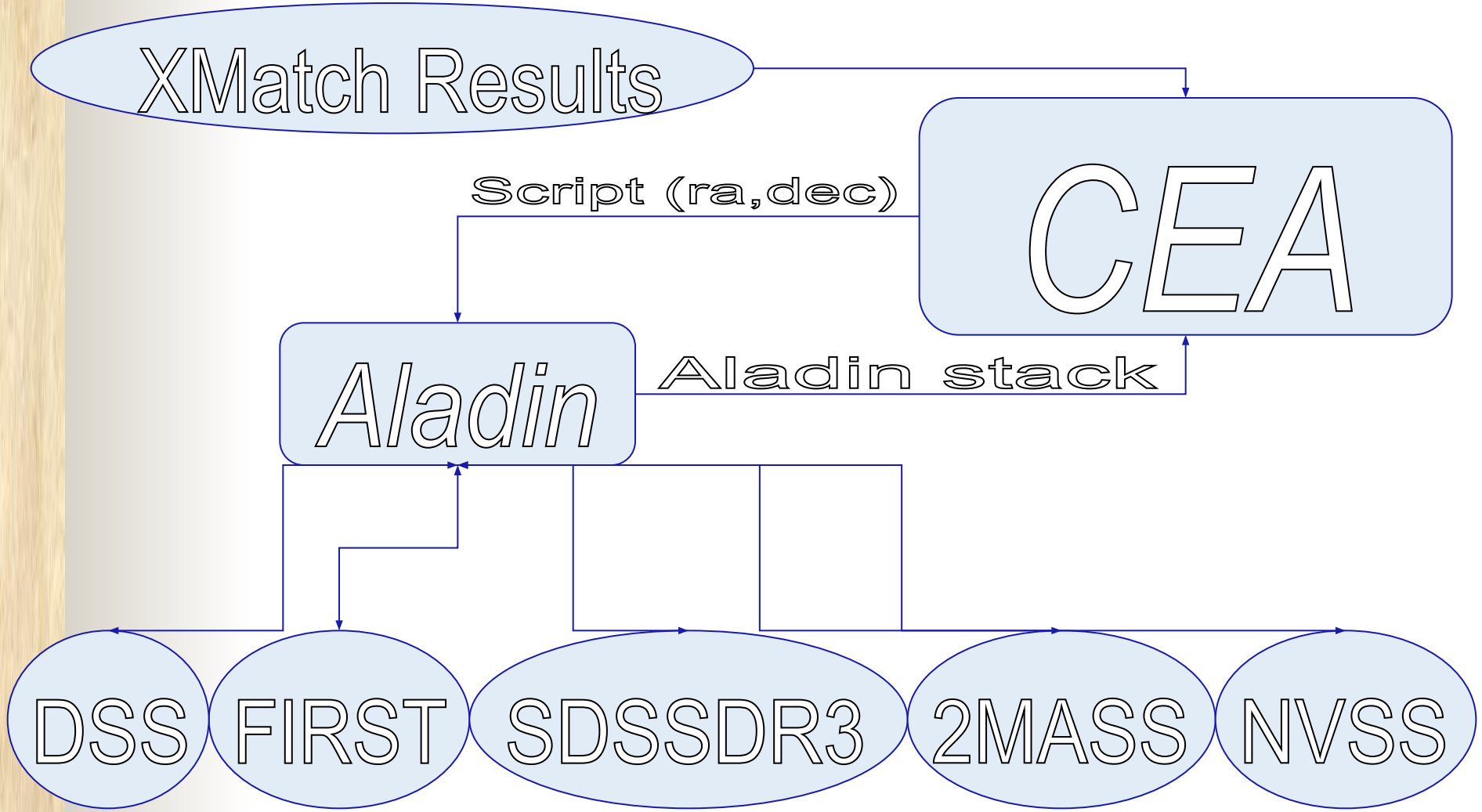
STEP1: QUERY TO MEDIATOR

```
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, deltaColorIndexURG])
& 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 > deltaColorIndexURG
```

STEP2: WORKFLOW SCRIPT

- Delete all null rows from Xmatch table
- Project Xmatch table (VOTable) into:
  - Ra coordinates list
  - Dec coordinates list
- Convert Ra and Dec coordinates from h:m:s format to format acceptable by Aladin

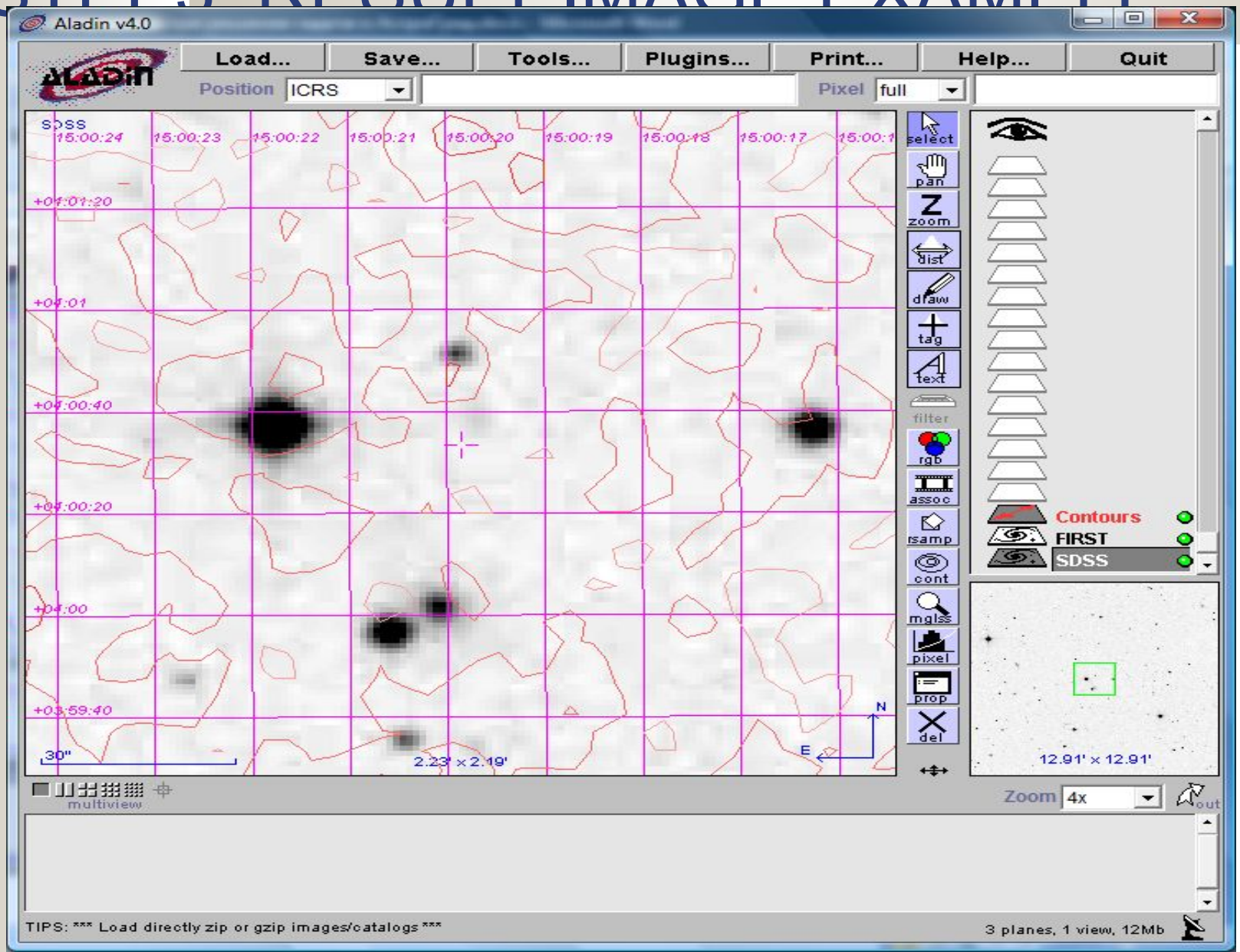
STEP3. DOWNLOADING IMAGES AND PROVIDING SUPERPOSITION OF THEM



STEP4. IMAGE ANALYSIS BY A SPECIALIST

- Launching Aladin (version 3.030_votech)
- Launching VO Desktop
- Opening Aladin stack, saved in VOSpace on a previous step
- Viewing, analyzing, probably editing obtained images using Aladin tools

STEP3: RESULT IMAGE EXAMPLE



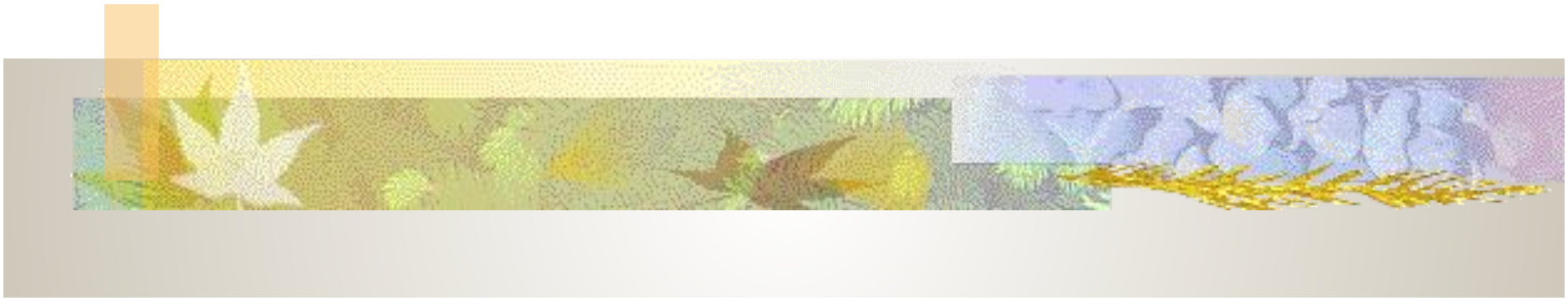
Principles of *application-driven mediation-based IS development*

A middleware for *application-driven mediation-based IS development* over multiple participating resources has been developed applying the following main principles:

- mediator specification **independence of the existing resources**;
- **canonical definitions** of the mediator specifications;
- **semantic integration of participating resources** in the canonical mediator specification;
- **integrated access to the participating [Grid-based] resources** registered at mediator applying the canonical model and mediator program rewriting system;
- **recursive structure of mediators**: each mediator can be registered as a resource in another mediator providing for cross domain interoperability;
- **mediator middleware independence** of particular [Grid] infrastructures and can be integrated with any of them.

A synergy of the mediation middleware and AstroGrid has been experienced for problem solving in the RVO environment.

Еще раз о проблематике семинара



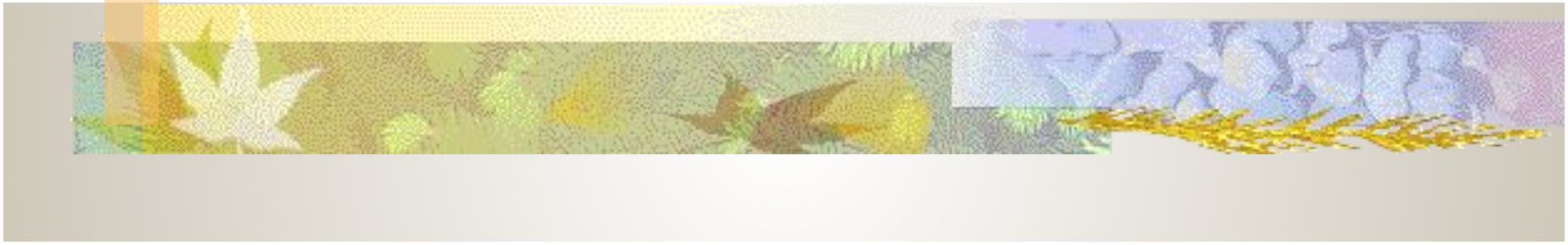
Краткая характеристика проблематики (ИТ)

- **Семантическая интероперабельность** информационных ресурсов и их интероперабельные **композиции**
- **Интеграция** множеств неоднородных информационных ресурсов
- **Типы, их композиции и операции** над коллекциями данных
- **Унификация** неоднородных информационных **моделей**
- **Композиции веб сервисов**
- **Потоки работ и их композиции**
- **Виртуальные организации** и их семантическая интероперабельность
- **Семантический Веб**.
- **Посредники в информационных системах** и их роль при интеграции информации и решении задач
- **Абстрактное описание предметных областей**
- **Спецификация задач над абстрактным описанием ПО**

Краткая характеристика проблематики (включая формальные методы)

- **Модели (языки) спецификаций предметных областей,** информационных ресурсов, задач. Спектр формализации моделей
- **Методы унификации** разнородных информационных моделей. **Канонические информационные модели**
- **Концептуальное моделирование** предметных областей и информационных ресурсов
- **Онтологическое моделирование**. Роль дескриптивных логик
- **Метод уточнения и** его роль в распределенных информационных системах. Инструменты на основе логики предикатов
- **Методы преобразования информационных моделей** в многоуровневых спецификациях. Методы архитектуры MDA
- **Методы переписывания программ** при интеграции информационных ресурсов

Литература



Список литературы

- 1. **Формальные методы и языки спецификации и проектирования ИС**

- Abrial J.-R. The B-Book. Assigning programs to meanings. Cambridge, University Press. 1996
- Abrial J.-R. B-Technology: Technical overview. B-Core (UK) Ltd., 1993
- Lano K. The B Language and Method: A Guide to Practical Formal Development. Springer-Verlag, 1996.
- Beckert B., Keller U., Schmitt P. Translating the Object Constraint Language into First-order Predicate Logic. Proc. VERIFY: Workshop at Federated Logic Conferences.
- Copenhagen, 2002.
- Ledang H., Souquieres J. Integration of UML and B Specification Techniques: Systematic Transformation from OCL Expressions into B. Proc. of APSEC 2002. Washington: IEEE Computer Society, 2002. P. 495-506.
- Morgan C. Programming from Specifications. Prentice Hall, 1994
- Back J.R., Von Right J. Refinement Calculus. N.Y.: Springer-Verlag, 1998.

Список литературы

- 1. **Формальные методы и языки спецификации и проектирования ИС (2)**
- Butler M. J. csp2B: A Practical Approach To Combining CSP and B. Formal Aspects of Computing. 2000. N. 12. P. 182-198.
- Хоар Ч. Взаимодействующие последовательные процессы. М.: Мир, 1989.
- Spivey J.M. The Z Notation (A reference manual). Prentice Hall, 1989
- Stepney S., Barden R., Cooper D. (Eds). Object Orientation in Z. Springer Verlag, WiC series, 1992
- R. Duke and G. Rose. Formal Object Oriented Specification Using Object-Z. Macmillan, 2000
- Kim S., Carrington D. A Formal Mapping between UML Models and Object-Z Specifications.
- ZB2000 Formal Specification and Development in Z and B: Proc. First International Conference of B and Z users. Springer-Verlag, 2000. P. 2-21
- Fitzgerald J., Larsen P.G., Mukherjee P., Plat N., Verhoef M. Validated Designs for Object-oriented Systems. Springer-Verlag, 2005.

Список литературы

- 2. **Введение в онтологическое моделирование**

- Kalinichenko L.A., Missikoff M., Schiappelli F., Skvortsov N. Ontological Modeling. Proc. of the Fifth Russian Conference on Digital Libraries RCDL'2003. St.-Petrsburg: St.-Petrsburg State University, 2003. -- P. 7 - 13.
- Ontolingua. http://ontolingua.stanford.edu
- OWL Web Ontology Language Guide, http://www.w3.org/TR/owl-guide/
- Kalinichenko L.A., Skvortsov N.A. Extensible ontological modeling framework for subject mediation In Proceedings of the 4-th Russian Scientific Conference "DIGITAL LIBRARIES: Advanced Methods and Technologies, Digital Collections, Oct. 15-17, 2002, Dubna
- Труды Симпозиума «Онтологическое моделирование», М:, ИПИ РАН, 2008

Список литературы

3. **Средства концептуального моделирования предметных областей**

- Калиниченко Л.А. СИНТЕЗ – язык определения, проектирования и программирования интероперабельных сред неоднородных информационных ресурсов, ИПИ РАН, 1993, 115 стр.
- Kalinichenko L.A. Compositional Specification Calculus for Information Systems Development Proceedings of the East-West Conference on Advances in Databases and Information Systems (ADBIS'99), Maribor, Slovenia, September 1999, Springer Verlag, LNCS
- Ceri S., Gottlob B., Tanca L. Logic programming and databases. Springer-Verlag, 1990
- Kalinichenko L.A., Stupnikov S.A., Martynov D.O. SYNTHESIS: a Language for Canonical Information Modeling and Mediator Definition for Problem Solving in Heterogeneous Information Resource Environments. Moscow: IPI RAN, 2007. - 171 p

Список литературы

4. **Вопросы композиционного проектирования систем**

- Брюхов Д.О. Конструирование информационных систем на основе интероперабельных сред информационных ресурсов: Дисс. канд. техн. наук: 05.13.11 -- М.: ИПИ РАН, 2003. 158 с.
- Ступников С.А. Автоматизация верификации уточнения при композиционном проектировании информационных систем и посредников. Системы и средства информатики: Спец. вып. Формальные методы и модели в композиционных инфраструктурах распределенных информационных систем. Под ред. И. А. Соколова. М.: ИПИ РАН, 2005. С. 96 -119

Список литературы

- 5. **Методы и средства конструирования канонических информационных моделей**

- Калиниченко Л.А. Методы и средства интеграции неоднородных баз данных. Москва, Наука, 1983, 420 стр.
- Kalinichenko L.A. Methods and tools for equivalent data model mapping construction Proc. EDBT'90 Conference. Springer-Verlag, 1990. 92-119
- L.A. Kalinichenko, N.A. Skvortsov Extensible ontological modeling framework for subject mediation. Proc. of the Fourth Russian Conference on Digital Libraries RCDL'2002. -- Dubna: JINR, 2002. V. 1. P. 99 - 119.
- Калиниченко Л.А., Ступников С.А., Земцов Н.А. Методы синтеза канонических моделей, предназначенных для достижения семантической интероперабельности неоднородных источников информации. ИПИ РАН, Москва, 2005

Список литературы

- 5. **Методы и средства конструирования канонических информационных моделей (2)**

- Формальные методы и модели в композиционных инфраструктурах распределенных информационных систем. Системы и средства информатики, специальный выпуск, ИПИ РАН, 2005 г., 304 стр.
- Ступников С.А. Формальная семантика ядра канонической объектной информационной модели. Системы и средства информатики: Спец. вып. Формальные методы и модели в композиционных инфраструктурах распределенных информационных систем. Под ред. И. А. Соколова. М.: ИПИ РАН, 2005. С. 40-68.
- Захаров В. Н., Калиниченко Л. А., Соколов И. А., Ступников С. А. Конструирование канонических информационных моделей для интегрированных информационных систем Информатика и ее применения. – М., – Т. 1, Вып. 2, 2007 г.

Список литературы

- 6. **Методы моделирования уточняющих спецификаций ИС**
- Ступников С.А. Моделирование композиционных уточняющих спецификаций. Дисс. канд. техн. наук: 05.13.17, М.: ИПИ РАН, 2005

- 7. **Архитектуры и методы спецификации предметных областей и решения задач над множеством неоднородных распределенных информационных ресурсов**

- Калиниченко Л.А. Методология организации решения задач над множественными распределенными неоднородными источниками информации. Сборник трудов Международной конференции «Современные информационные технологии и ИТ-образование». - М.: МГУ, 2005. - С. 20 – 37
- Briukhov D.O., Kalinichenko L.A., Skvortsov N.A. Information sources registration at a subject mediator as compositional development Advances in Databases and Information Systems: Proc. of the 5th East European Conference. Berlin-Heidelberg: Springer-Verlag, 2001. P. 70-83
- Stupnikov S.A., Kalinichenko L.A., Bressan S. Interactive discovery and composition of complex Web services. In Proceedings of the East-European Conference on “Advances in Databases and Information Systems" (ADBIS'06), Thessaloniki, Springer, LNCS, 2006.

Список литературы

- 7. **Архитектуры и методы спецификации предметных областей и решения задач над множеством неоднородных распределенных информационных ресурсов (2)**

- Д.О. Брюхов, А.Е. Вовченко, В.Н. Захаров, О.П. Желенкова, Л.А. Калиниченко, Д.О. Мартынов, Н.А. Скворцов, С.А.Ступников. Архитектура промежуточного слоя предметных посредников для решения задач над множеством интегрируемых неоднородных распределенных информационных ресурсов в гибридной грид-инфраструктуре виртуальных обсерваторий. Информатика и ее применения. – М., – Т. 2, Вып. 1, 2008 г.

- 8. **Дополнительная литература**

- Borger E., Stark R. Abstract State Machines: A Method for High-Level System Design and Analysis. Springer Verlag, 2003.
- OMG Unified Modeling Language Specification http://www.omg.org
- Bruel J.M., France R.B. Transforming UML models to formal specifications UML'98: Beyond the Notation: Proc. of 1st International Workshop. Springer-Verlag, 1999

IPI RAN Laboratory for compositional information systems development

Various forms of compositions are studied, including :

- Interoperable compositions of pre-existing components for IS design;
- Compositions of heterogeneous information collections (including integration);
- Web services and Workflow compositions;
- Type compositions in database operations over object collections;
- Subject mediators development and compositions.

Web site of the group: http://www.ipi.ac.ru/synthesis/ (only publications page is updated regularly)

Conferences and groups organized by the laboratory initiative for consolidation of database and information systems community:

- Moscow ACM SIGMOD Chapter
- East European Conference ADBIS (11^{th} conference is planned in September)
- Russian Conference on Digital Libraries RCDL (9^{th} conference is planned in October)

Collaborative projects of the Laboratory

- **Object model integration**, project with **GTE Labs** (Michael Brodie), 1992 – 1993
- **Semantically Interoperable Information Systems**, **INTAS Project, GMD IPSI** (Coordinator: E.Neuhold; W.Klas, P.Fankhauser, K. Aberer), **PRISM laboratory**, France (George Gardarin), **Kiev State University** (N. Nikitchenko), 1995 – 1997
- **Compositional software development** using pre-existing heterogeneous information resource. Research project with the **Siemens Software Arch. and Reuse Department**, 1994 - 1997
- **Modelling and management of semi-structured data for dynamic world-wide-web applications**, **INTAS Project**, **Aristotle University** of Thessaloniki (Coordinator: Y.Manolopoulos), **Rome University** III (P. Atzeni, G. Mecca, R. Torlone), St. Petersburg University (B. Novikov), **Yerevan University** (M. Manukyan), 2000 – 2001
- **DELOS Network of Excellence**, 2000 – 2003; DELOS WG on Ontology Harmonization (Martin Doerr); DELOS WG on Metadata Registries (Thomas Baker)
- **UNESCO project: Digital Libraries in Education**, groups from USA and Europe coordinated by L.A.Kalinichenko, 2002
- Various research projects granted by the Russian Foundations for Basic Research (**RFBR**) and **RAS**
- **Details could be accessed at www.ipi.ac.ru**