Поддержка повторного использования спецификаций потоков работ за счет обеспечения их независимости от конкретных коллекций данных и сервисов

© Брюхов Д.О. © Вовченко А.Е. © Калиниченко Л.А.
ИПИ РАН,
Москва
brd@ipi.ac.ru itsnein@gmail.com leonidk@synth.ipi.ac.ru

Аннотация

Статья рассматривает вопросы организации исследований в науках с интенсивным использованием данных (НИИД). Конкретно в ней изучается проблема повторного использования потоков работ в научных исследованиях. В статье представлен подход к встраиванию предметных посредников в среду для совместных исследований в НИИД. Этот подход позволяет создавать методы и алгоритмы решения задач независимо от конкретных реализаций ресурсов (данных и сервисов). За счет обеспечения независимости потоков работ от конкретных коллекций данных и сервисов существенно упрощается возможность повторного использования потоков работ.

1 Введение

Науки с интенсивным использованием данных (НИИД) развиваются в рамках новой парадигмы научных исследований (так называемой 4-й парадигмы [14]), согласно которой новые знания образуются в результате анализа разнообразных данных, накопленных в результате проведения измерений, наблюдений, моделирования, вычислений. Формулирование этой парадигмы явилось результатом осознания все возрастающей роли данных для развития науки, научных открытий практически во всех научных областях. Данные становятся ключевым источником получения знаний в НИИД. При этом объем, разнообразие и качество накапливаемых данных быстро растет отчасти благодаря быстрому развитию техники наблюдений и измерений различных природных явлений и процессов, введению в практику новых методов и инструментов наблюдения. Поэтому системы с интенсивным использованием данных имеют существенное пересечение с быстро развиваемой областью, именуемой «Big Data».

Вместе с тем, в НИИД «ученые, вместо того, чтобы заниматься исследованиями, затрачивают большую часть своего времени на поиск данных, манипулирование, обмен данными. И такое положение все время усугубляется» (наблюдение DoE Office of Science Data Management Challenge в USA).

Наиболее заметны следующие проблемы организации исследований в НИИД:

1) Создаваемые в НИИД методы анализа данных и алгоритмы решения задач как правило ориентированы на конкретные коллекции данных, находящиеся в поле зрения конкретных ученых в конкретный момент. Из-за этого отсутствует возможность повторного использования таких методов, алгоритмов и их реализаций над другими данными, в других коллективах НИИД.

2) Отсутствует практика накопления и повторного использования методов анализа данных, алгоритмов решения задач и их реализаций в научном сообществе НИИД. Фактически опыт проведения исследований, методы решения задач анализа данных в НИИД не накапливаются.

3) В НИИД отсутствует практика формирования ИТ-базированных, согласованных в сообществах концептуальных определений научных областей (включающих их структуру, понятия, спецификации методов, задач, техник проведения измерений и экспериментов, и пр.).

Данная статья подготовлена в рамках проекта[1], ориентированного на преодоление названных проблем. Для преодоления проблемы (2) предлагается использовать потоки работ как

Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2013, Ярославль, Россия, 14-17 октября 2013 г.

[1] Проект “Обеспечение повторного использования реализаций методов анализа информации и алгоритмов решения задач в научных областях с интенсивным использованием данных" в рамках программы фундаментальных исследований Президиума РАН № 16 «Фундаментальные проблемы системного программирования»

универсальное средство определения и реализации методов анализа данных, алгоритмов решения задач и их композиций. Опыт проведения исследований с интенсивным использованием данных в научном сообществе НИИД предлагается накапливать в виде потоков работ и их метаописаний. Средства накопления спецификаций потоков работ реализованы при этом на основе обоснованного выбора одного из существующих международных проектов подобных систем (таких как myExperiment [4], Wf4Ever [11], VisTrails [10], Trident [9], и др.). Одним из существенных недостатков таких проектов является отсутствие возможности использования в них концептуальных определений коллекций данных, обрабатываемых потоками работ (проблема 3), и, как следствие этого, ориентированность потоков работ на конкретные коллекции данных, что препятствует возможности повторного использования спецификаций потоков работ и их реализаций над другими данными в других исследованиях НИИД (проблема 1). В статье показано, как преодолеть названные недостатки за счет введения концептуальных спецификаций в практику определения потоков работ и задания отображений в них конкретных коллекций данных на основе техники предметных посредников. Тем самым удается обеспечить независимость накапливаемых для повторного использования спецификаций потоков работ от конкретных коллекций данных, а также при необходимости применить интеграцию конкретных коллекций данных для образования адекватных концептуальных коллекций.

2 Среды для публикации и повторного использования потоков работ

В настоящем разделе дан краткий обзор систем, обеспечивающих публикацию и повторное использование спецификаций потоков работ.

Особо стоит выделить среду для совместных исследований myExperiment [4], в которой ученые могут публиковать потоки работ для решения задач. Среда myExperiment была введена в 2007 году и в настоящее время является одной из самых больших репозиториев потоков работ (в ней содержится более 2000 потоков работ), используется тысячами ученых в различных областях науки. Среда myExperiment позволяет публиковать потоки работ в различных системах управления потоками работ. Для ряда систем управления потоками работ (таких, как Taverna [6], Galaxy [8], Trident [9]) поддерживаются дополнительные возможности такие, как управление метаданными, извлечение информации об используемых сервисах, визуализация потоков работ.

Другим примером репозитория потоков работ является проект ER-flow [5] (проект FP7 "Building a European Research Community through Interoperable Workflows and Data"), являющийся продолжением проекта SHIWA. Проект ER-flow предоставляет ученым программную поддержку для создания, обмена и запуска потоков работ в различных системах управления потоками работ (ASKALON, Galaxy, GWES, Kepler, LONI Pipeline, MOTEUR, Pegasus, P-GRADE, ProActive, Triana, Taverna, WS-PGRADE).

Системы управления потоками работ в науке поддерживают доступ к широкому набору уже существующих баз данных и сервисов анализа данных в различных областях науки (в биологии, астрономии, социальных науках, и др.), использование которых позволяет упростить процесс создания потоков работ.

Репозитории потоков работ позволяют ученым находить интересующие их потоки работ, воспроизводить результаты этих потоков работ, повторно использовать существующие потоки работ для решения задач в рамках названных выше ограничений.

Для конкретизации рассмотрения в данной статье предполагается использовать myExperiment с ориентацией на систему управления потоками работ Taverna [6]. Taverna – это система управления потоками работ, которая может быть использована в различных областях науки. Она предоставляет набор сервисов для создания и выполнения разнообразных потоков работ. Taverna была создана в рамках проекта myGrid [7].

3 Проблемы повторного использования потоков работ

Taverna предоставляет средства для поиска (по тегам) потоков работ в среде myExperiment. Найденные потоки работ можно запускать как с исходными значениями входных параметров, предоставленными разработчиками, так и с произвольными значениями. Это позволяет воспроизвести результаты исследования других ученых с целью возможного повторного использования разработанных потоков работ. Тем не менее зачастую повторное использование может оказаться невозможным.

Спецификация потока работ в Taverna задается в виде направленного графа. Потоки работ в Taverna реализуют модель потоков данных (data flow model). Таким образом,.поток работ состоит из сервисов, представляющих собой программные компоненты (такие как веб-сервисы), и направленных связей между ними, выражающих зависимости по данным. Taverna поддерживает широкий набор как локальных, так и удаленных сервисов в различных областях науки. В частности, Taverna обеспечивает доступ к произвольным WSDL и REST сервисам; к конкретным веб сервисам, таким как BioMoby [15], BioMart [12] и SoapLab [16]; к локальным Java сервисам (BeanShell скрипты); к базам данных посредством JDBC. Taverna поддерживает использование вложенных потоков работ. Это позволяет встраивать уже существующие потоки

работ (возможно разработанные другими учеными) при создании новых потоков работ.

Одной из главных проблем повторного использования потоков работ в Taverna является зависимость спецификаций потоков работ от конкретных коллекций данных и/или сервисов. В Taverna каждый сервис настраивается на доступ к конкретным сервисам и базам данных. Это не позволяет повторно использовать такие потоки работ, если необходимо, например, обрабатывать другие коллекции данных. Также, если какой-либо из сервисов или база данных в настоящий момент недоступны, то весь поток работ не сможет быть выполнен.

Данная статья нацелена прежде всего на решение проблемы повторного использования потоков работ в Taverna над базами данных. Taverna поддерживает ряд способов доступа к базам данных из потока работ:

1. Создание веб сервиса, реализующего доступ к базе данных. Доступ к этому веб сервису из потока работ осуществляется по протоколу SOAP;
2. Полная реализация интерфейса расширения (extension point) Taverna, включающего поддержку языка запросов к базе данных и графический интерфейс для конструирования запросов и предоставления пользователю метаданных подключаемой базы данных. В Taverna этот подход реализован для сервиса BioMart [12] и в плагине AstroTaverna [13];
3. Использование существующих сервисов BioMart для доступа к подключаемой базе данных;
4. Использование JDBC сервиса для доступа к базам данных.

Возможность подключения нового ресурса через BioMart заслуживает отдельного рассмотрения. BioMart (а точнее BioMart портал) представляет собой систему управления данными, ориентированную на выполнение разнообразных запросов над биологическими данными. В портале системы можно найти нужные ресурсы по метаданным, а также задать к ним запрос и получить результат. Также запросы могут быть заданы над несколькими конкретными базами данных, зарегистрированными в портале. Данные из BioMart могут быть получены посредством веб страницы, графического или консольного инструментария, или из программ посредством веб-сервисов либо напрямую через perl или java АПИ.

С другой стороны, BioMart (а точнее BioMart сервис) представляет собой адаптер, унифицирующий интерфейс различных баз данных, таких как MS SQL Server, PostgreSQL, MySQL, DB2, Oracle. По сути, любая (из поддерживаемых) база данных может быть оформлена как BioMart сервис, после чего полученный сервис подключается к порталу. С точки зрения схемы ресурса, при создании BioMart сервиса возможно определение взглядов (SQL views) над исходной схемой для ее модификации (удалить атрибуты, убрать какие-то таблицы, добавить ключи, и др.). Также, для повышения производительности взгляды можно материализовать. BioMart автоматически обновляет материализованные взгляды в случае изменения исходных данных в ресурсе. Кроме того, можно устанавливать связи между различными базами данных (по ключам), образуя их федерацию.

С концептуальной точки зрения схемы BioMart сервисов определяются на основе схем ресурсов. Это подход известен в литературе как GAV [2] и обладает рядом недостатков, основным из которых является слабая масштабируемость, т.к. добавление (удаление) одного из ресурсов влечет за собой изменение федеративной схемы. Инструментарий Taverna предоставляет доступ не к BioMart порталу, а к отдельным BioMart сервисам. Чтобы добавить новую операцию в поток работ, выбирается конкретный BioMart сервис, с конкретной схемой, и формулируется конкретный запрос, что также затрудняет повторное использование потока этого работ.

Основное отличие предлагаемого в настоящей работе подхода заключается в поддержке концептуальной схемы предметной области для спецификации потоков работ и введении промежуточного слоя предметных посредников, обеспечивающего отображение схем произвольных конкретных ресурсов (баз данных и сервисов) в концептуальную схему, интеграцию ресурсов. Благодаря этому спецификация потоков работ не требует изменения при изменении ресурсов, что является необходимым условием обеспечения повторного использования потоков работ.

4 Инфраструктура предметных посредников как средство решения проблем повторного использования

4.1 Концепции инфраструктур предметных посредников

Основной идеей инфраструктуры решения задач над неоднородными информационным ресурсами является введение промежуточного слоя между ресурсами и потребителями информации, образуемого предметными посредниками [1]. Каждый предметный посредник поддерживает спецификацию предметной области для решения некоторого класса задач.

Посредники реализуют подход к решению задач, ориентированный на проблему. В рамках подхода, ориентированного на проблему (подхода, «движимого приложением»), формулируется концептуальная спецификация задачи, включающая базовые сущности и понятия предметной области, функции, процессы и пр. Такое определение предметной области, представляет собой спецификацию предметного посредника для решения класса задач. Сущности и понятия предметной области, определенные таким образом, не зависят от существующих информационных

ресурсов. В терминах предметной области формулируются программы для решения задачи на языке правил посредника и на языках программирования. Для решения конкретной задачи выявляются инфраструктуры, содержащие ресурсы, необходимые для ее решения (например, гриды, облачные инфраструктуры, репозитории данных, и др.). Далее, идентифицируются ресурсы, релевантные задаче, используя реестры доступных инфраструктур. Релевантные задаче ресурсы регистрируются в предметных посредниках, задающих отображение схем ресурсов в концептуальную спецификацию.

Таким образом, при изменении набора ресурсов, спецификация алгоритма решения задачи остается неизменной, и может быть повторно использована на другом наборе коллекций данных.

4.2 Обеспечению независимости потоков работ от данных на основе предметных посредников

Как было отмечено выше, все сервисы в потоках работ Taverna определены в терминах конкретных сервисов и баз данных, что не позволяет задавать спецификации потоков работ независимо от конкретных ресурсов.

По сути, посредники представляют собой виртуальные базы данных, и в потоках работ Taverna их можно подключать аналогично обычным базам данных. Возможны 2 способа подключения посредников к Taverna: посредством веб сервиса и посредством разработанного плагина (соответствующие 1-му и 2-му способам, рассмотренным в разделе 3). При первом способе над посредником создается веб сервис, реализующий интерфейс посредника. Доступ к посреднику из потоков работ Taverna осуществляется посредством этого веб сервиса по протоколу SOAP. Вторым способом подключения предметных посредников к Taverna может являться разработка специального плагина под средство разработки потоков работ Taverna Workbench. Taverna предоставляет возможность создания подобных плагинов, посредством интерфейса расширения (extension point), для добавления и расширения функциональности Taverna Workbench. Этот плагин сможет предоставлять графический интерфейс для помощи в конструировании запросов к предметным посредникам и интерфейс для доступа к метаданным предметного посредника.

Все доступные в Taverna ресурсы, используемые в качестве узлов в потоках работ, могут быть использованы также посредством посредников. В частности, предметные посредники поддерживают использование WSDL сервисов в виде функций. Конкретные веб-сервисы (например, BioMoby, BioMart и SoapLab) также могут быть использованы из посредника. BeanShell скрипты могут быть оформлены в виде программ на Java над предметным посредником, либо в виде функции предметного посредника. Базы данных подключаются к посреднику посредством адаптеров.

Концептуальные коллекции с технической точки зрения могут быть использованы точно также как обычные базы данных в Taverna. С помощью предметных посредников в виде концептуальных коллекций могут быть оформлены любые базы данных. Главное отличие концептуальных коллекций от обычных заключается в том, что их схема остается неизменной независимо от набора фактически используемых ресурсов. В результате, запросы к концептуальной коллекции, и следовательно, поток работ остаются неизменными при изменении набора конкретных ресурсов. Таким образом может быть получена спецификация потока работ, определяемая в терминах предметной области предметного посредника и не зависящая от конкретных ресурсов. Это решает одну из основных проблем повторного использования потоков работ.

5 Пример применения подхода к обеспечению независимости спецификации потоков работ на основе задачи определении вторичных стандартов

В этом разделе мы рассмотрим предлагаемый нами подход на задаче определения вторичных стандартов для фотометрической калибровки оптических компонентов космических гамма-всплесков [3], поставленной Институтом Космических Исследований РАН. Задача заключается в том, что по координатам площадки, требуется найти в ней звезды, удовлетворяющие ряду условий (не переменные, точечные, с хорошими изученными параметры). Такие звезды называется «стандартами» и могут быть использованы для калибровки новых поступающих данных.

5.1 Описание схемы посредника для задачи определения вторичных стандартов

На Рис. 1 представлена схема посредника, разработанная для решения этой задачи. Она включает в себя описание концептов, необходимых для решения задачи, таких как: экваториальные координаты (CoordEQJ); фотометрическую систему (PhotometricSystem); фотометрическую полосу (Passband); магнитуду в некоторой фотометрической системе (Magnitude); абстрактный астрономический объект (Astronomical Object); звезду (Star); стандарт (Standard); изображение (Image). Также схема посредника содержит функции, необходимые для решения задачи, включая: метод кросс-идентификации (matchObjects); метод вычисления цветового индекса (colorIndex); метод проверки типа объекта по некоторому эталонному каталогу (каталогам) (checkType); метод проверки, является ли звезда переменной на основе данных из многих других ресурсов (isVariable).

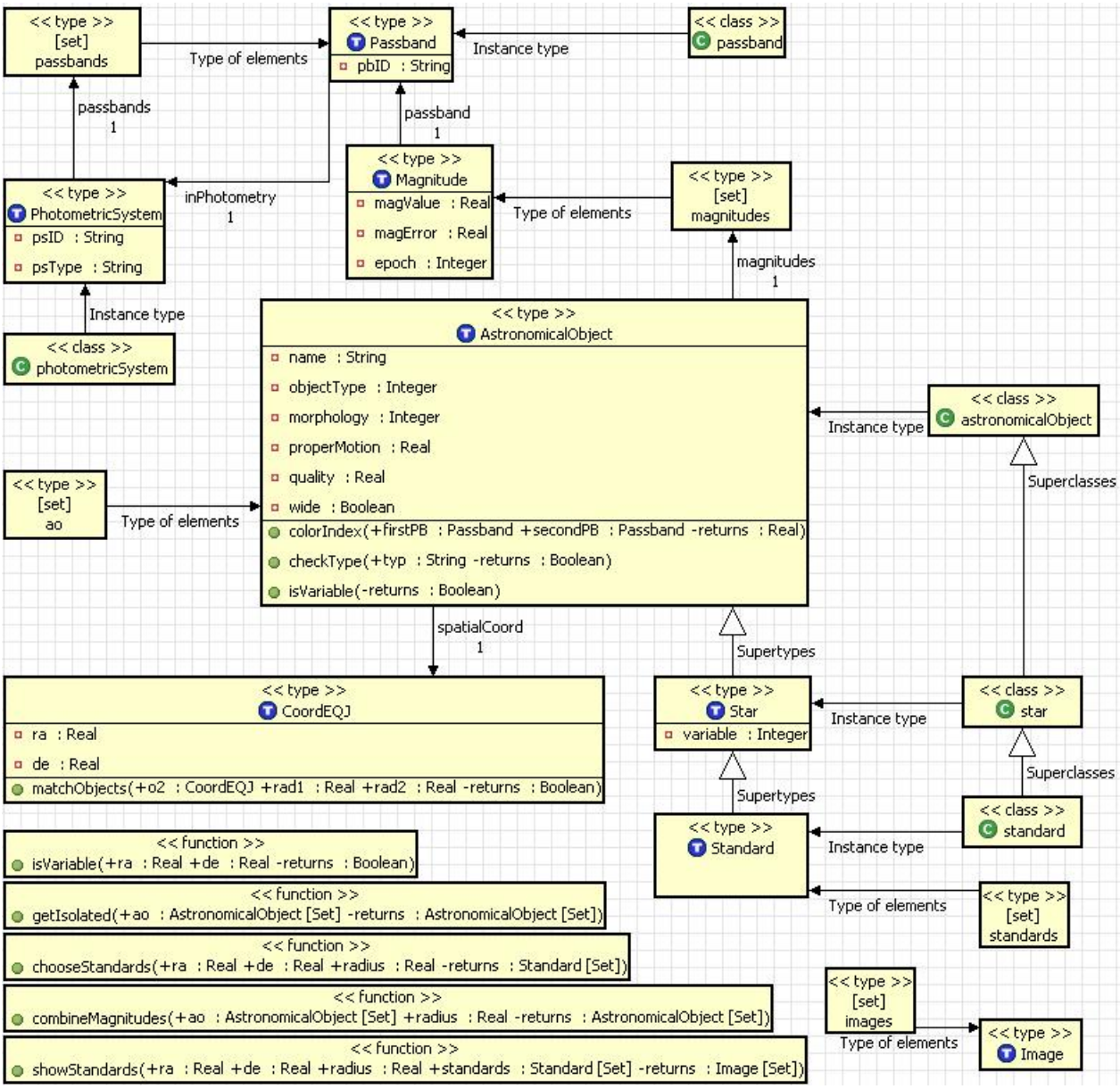

Рис. 1 Схема посредника для задачи определения вторичных стандартов

Представленная схема не зависит от конкретных ресурсов, используемых для решения задачи. Каталоги SDSS, USNOB-1, 2MASS, GSC, UCAC – основные ресурсы, используемые для извлечения стандартов. Именно среди этих каталогов отбираются все звезды, удовлетворяющие параметрам. Каталоги VSX, ASAS, GCVS, NSVS используются для проверки факта переменности выбранных стандартов. Список ресурсов может со временем меняться, но при этом схема посредника останется неизменной и методы решения задач определения вторичных стандартов также останутся неизменными.

5.2 Программа решения задачи определения вторичных стандартов

Задача определения стандартов была сформулирована в виде программы (последовательности правил) над схемой, рассмотренной выше. Параметром программы является площадка на небесной сфере, в которой произошел гамма-всплеск. Площадка характеризуется центром с координатами queryRA, queryDE и радиусом radius. Программа посредника состоит из восьми последовательных правил.

Правило 1 – В первом правиле среди всех астрономических объектов выбираются те, что попадают в указанную площадку. При этом нас интересуют только координаты (ra, de), звездные величины в различных полосах (magnitudes), тип объекта (objectType), собственное движение (properMotion) и качество данных (quality). Это правило на языке правил посредников (язык СИНТЕЗ [17]) выглядит следующим образом:

```
r(x/[ra, de, name, magnitudes, objectType,
properMotion, quality])
:- astronomicalObject(x1/[ra: spatialCoord.ra, de:
spatialCoord.de, name, objectType, properMotion,
quality, magnitudes])
& ra < queryRA + radius & ra > queryRA - radius
```

```
& de < queryDE + radius & de > queryDE - radius
```

Правило продуцирует коллекцию r, состоящую из астрономических объектов (astronomicalObject), содержащих необходимые атрибуты и удовлетворяющих ограничениям на координаты, указанные в теле правила.

Правило 2 – Во втором правиле отсеиваются неизолированные объекты. Изолированные объекты – это объекты, в некоторой окрестности которых на небесной сфере не наблюдается других объектов:

```
getIsolated(r1, r2);
```

Правило 3 – В третьем правиле среди ранее выбранных объектов отсеиваются галактики, и выбираются звезды с очень малым собственным движением и качественными фотометрическими данными:

```
r3(x/[ra, de, name, magnitudes])
:- r2(x1/[ra, de, name, objectType, properMotion,
quality, magnitudes])
& checkType(ra, de, 'Galaxy', nType) & nType = false
& objectType = Star
& properMotion < 0.01
& quality  < 0.01
```

Правило 4 - В четвертом правиле используются объекты, полученные в первом правиле. Среди объектов этого класса выбираются только те, для которых верно, что они переменные. Переменность определяется с помощью функции isVariableByMagnitude.

```
r4(x/[ra, de, name])
:- r1(x1/[ra, de, name, magnitudes])
& isVariablebyMagnitudes(ra, de, isVar) & isVar = true
```

Правило 5 - В пятом правиле выбираются переменные звезды из каталогов переменных звезд: GCVS, VSX, NSVS, ASAS.

```
r4(x/[ra, de, name])
:- variableStar(x1/[ra: spatialCoord.ra, de:
spatialCoord.de, name])
```

Правило 6 - В шестом правиле, производится кросс-идентификация объектов из класса кандидатов в стандарты (результат правила 3), и класса переменных звезд, посредством вызова функции xmatch.

```
xmatch(r3, r4, r5);
```

Правило 7 - В седьмом правиле из класса кандидатов в стандарты, полученного после кросс-идентификации, выбираются только те объекты, для которых не нашлось близко расположенного переменного объекта (distance > 0.01). На практике, это означает что кандидат в стандарты – не переменный объект.

```
r6(x/[ra, de, name magnitudes])
:- r5(x1/[ra, de, name, magnitudes, distance])
& distance > 0.01
```

Правило 8 – В предыдущем правиле построена коллекция r6, содержащая стандартные звезды. В заключительном правиле стандарты маркируются на изображение площадки гамма-всплеска, и предоставляются пользователю для утверждения.

```
r7(im/Image)
:- r6(x/ra, de, name, magnitudes])
& showStadards(ra, de, radius, magnitudes, im)
```

5.3 Описание Веб сервиса для доступа к посреднику для задачи определения вторичных стандартов

Для доступа к предметному посреднику решения задачи определения стандартов был разработан Веб сервис. Этот веб сервис включает в себя следующие методы, реализующие описанные выше правила:

executeQuery – выполняет правило посредника [17]. Этим правилом достаются кандидаты в стандарты. В качестве правила используется комбинация из описанных выше правил 1-3 (раздел 5.2). Данные возвращаются в формате SynthClass[2].

getVariableStarsFromCatalogues - получает из посредника коллекцию переменных звезд в заданной области из каталогов переменных звезд (правило 5). Данные возвращаются в формате SynthClass.

getVariableStarsByMagnitudes - получает из посредника коллекцию переменных звезд в заданной области, определяя переменная ли она по магнитудам (правило 1 и 4). Данные возвращаются в формате SynthClass.

removeVariableStars - получает коллекцию стандартов, и коллекцию переменных (аналог правил 6 и 7 реализованных одной функцией). Из первой удаляются те объекты, которые содержатся во второй.

removeStarsWithAnomalyMagnitudes - отсеивает аномальные звезды из входной коллекции объектов. Это дополнительный метод, не описанный выше в правилах. Был добавлен по настоянию астрономов для обеспечения большей точности результата.

getAladinCandidates – по полученной коллекции объектов возвращает изображение (аналоги правила 8), которое может быть открыто специалистом из программы Aladin [19], популярной среди астрономов.

[2] Формат представляет собой расширение стандартного для виртуальной обсерватории представления таблиц VOTable [18]. Расширения обеспечивают возможность представления коллекций объектов сложной структуры.

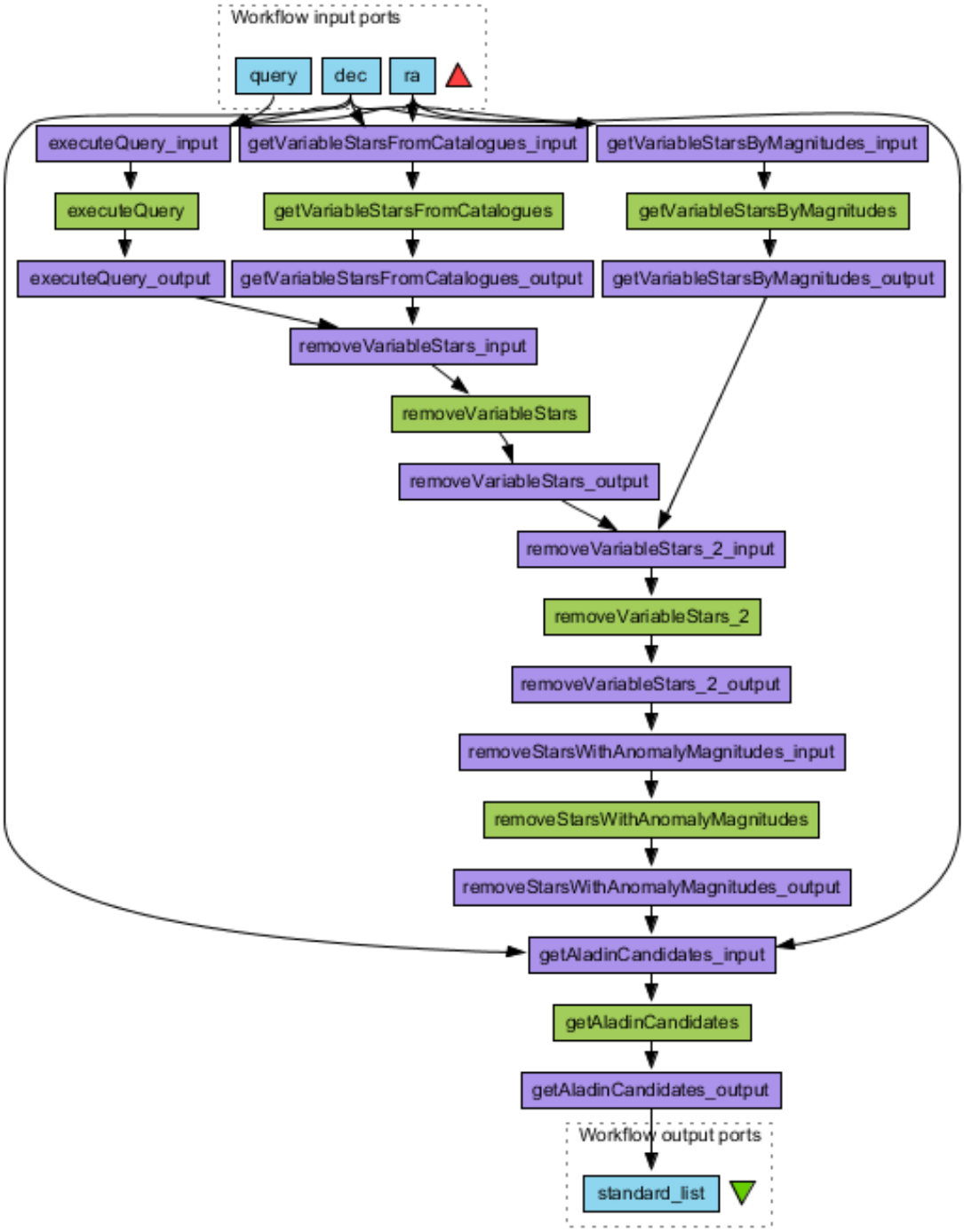

Рис. 2 Поток работ решения задачи вторичных стандартов в среде Taverna

5.4 Описание потока работ решения задачи определения вторичных стандартов в среде Taverna

На Рис. 2 представлен поток работ решения задачи вторичных стандартов в среде Taverna. Входными параметрами его являются координаты площадки на небесной сфере, в которой произошел гамма-всплеск.

Поток работ представляет собой набор вызовов методов Веб сервиса, описанного выше. Также в потоке работ присутствуют вспомогательные функции преобразования входных и выходных параметров методов в формат XML.

Результатом выполнения этого потока работ является изображение Aladin [19] с наложенным на него списком стандартов. На Рис. 3 показан пример результата, получаемого специалистом. Результат включает в себя изображение, а также отмеченные на изображении объекты – кандидаты в стандарты, удовлетворяющие всем требованиям.

5 Заключение

Предлагаемый подход по встраиванию предметных посредников в среду организации исследований в НИИД позволяет упростить

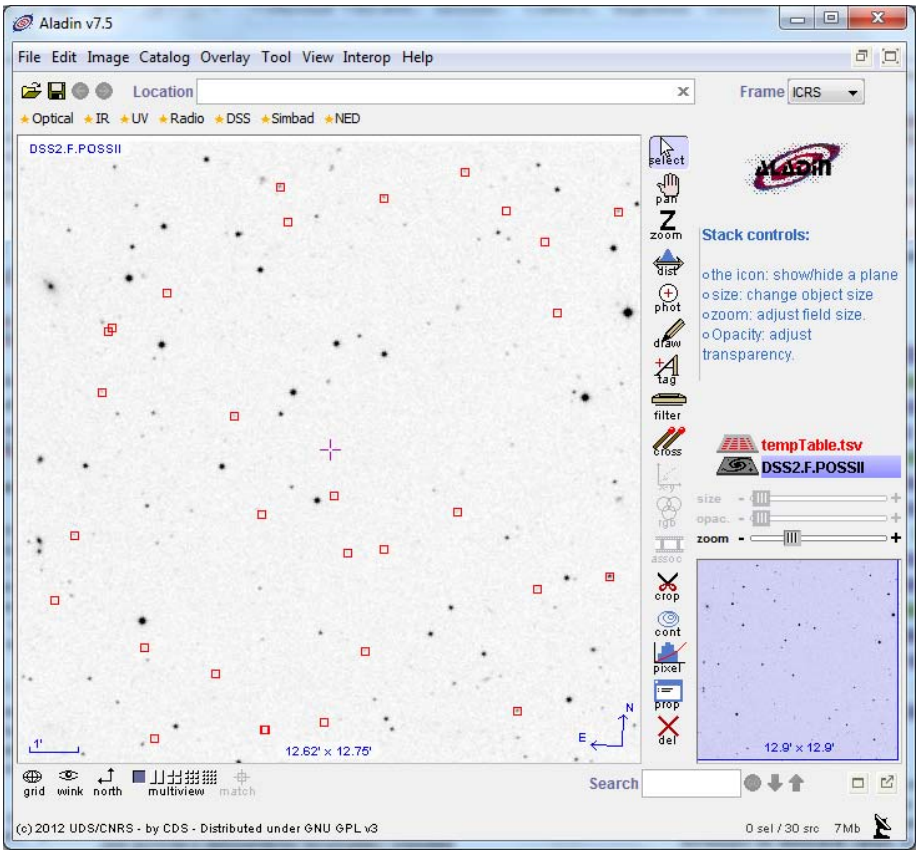

Рис. 3 Изображение найденных кандидатов в стандарты

решение ряда проблем таких, как: накопление методов анализа данных, алгоритмов решения задач и их реализаций в научном сообществе; воспроизведение и повторное использование таких алгоритмов и методов; формирование ИТ-базированных концептуальных определений научных областей; использование методов и средств высокоуровневых декларативных определений методов анализа данных и алгоритмов решения задач в НИИД. Хотя статья рассматривает предлагаемый подход применительно к конкретной среде myExperiment и системе управления потоками работ Taverna, предлагаемый подход может быть аналогично использован в других средах с другими системами управления потоками работ.

Литература

[1] Брюхов Д.О., Вовченко А. Е., Захаров В.Н., Желенкова О.П., Калиниченко Л.А., Мартынов Д.О., Скворцов Н.А., Ступников С.А. 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., 2008. – Т. 2, Вып. 1. – С. 2-34.

[2] Alon Y. Halevy. Answering Queries Using Views: A Survey. VLDB Journal, 10(4), 2001.

[3] Вовченко А.Е., Вольнова А.А., Денисенко Д.В., Калиниченко Л.А., Куприянов В.В., Позаненко А.С., Скворцов Н.А., Ступников С.А. Применение средств виртуальной обсерватории для выбора вторичных стандартов поля при фотометрии оптического послесвечения гамма-всплесков // Труды Всероссийской астрономической конференции ВАК-2010 «От эпохи Галилея до наших дней». – САО РАН: Нижний Архыз. – 2010.

[4] De Roure, D., Goble, C. and Stevens, R. (2009) The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows. Future Generation Computer Systems 25, pp. 561-567

[5] Mark Santcroos. Experiences from workflow sharing using the SHIWA Workflow Repository for application porting to DCI. EGI Community Forum Book of Abstracts, EGI, Manchester, UK, 2013.

[6] Katherine Wolstencroft, Robert Haines, Donal Fellows, Alan Williams, David Withers, Stuart Owen, Stian Soiland-Reyes, Ian Dunlop, Aleksandra Nenadic, Paul Fisher, Jiten Bhagat, Khalid Belhajjame, Finn Bacall, Alex Hardisty, Abraham Nieva de la Hidalga, Maria P. Balcazar Vargas, Shoaib Sufi, and Carole Goble. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. Nucleic Acids Research, First published online May 2, 2013.

[7] myGrid project http://www.mygrid.org.uk/

[8] Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010 Aug 25;11(8):R86.

[9] Roger Barga, Jared Jackson, Nelson Araujo, Dean Guo, Nitin Gautam, Yogesh Simmhan. The Trident Scientific Workflow Workbench. Proceeding of the 2008 Fourth IEEE International Conference on eScience, Pages 317-318, December 07-12, 2008.

[10] Steven P. Callahan, Juliana Freire, Emanuele Santos, Carlos E. Scheidegger, Claudio T. Silva and Huy T. Vo. VisTrails: Visualization meets Data Management. Proceedings of ACM SIGMOD 2006.

[11] Wf4Ever project http://www.wf4ever-project.org/

[12] Kasprzyk A. BioMart: driving a paradigm change in biological data management. Database (Oxford) 2011.

[13] Walton N. A., Witherwick D. K., Oinn T., Benson K. M. Taverna and workflows in the virtual observatory, Astronomical Data Analysis Software and Systems ASP Conference Series, Vol. 394, Proceedings of the conference held 23-26 September, 2007, p 309.

[14] The Fourth Paradigm: Data-Intensive Scientific Discovery. Tony Hey, Stewart Tansley, and Kristin Tolle, Eds. Microsoft Research, Redmond, WA, 2009. 286 pp.

[15] M. D. Wilkinson, D. Gessler, A. Farmer, L. Stein. The BioMOBY Project Explores Open-Source, Simple, Extensible Protocols for Enabling Biological Database Interoperability. In Proceedings of the Virtual Conference on Genomics and Bioinformatics (2003).

[16] Martin Senger, Peter Rice, Tom Oinn. Soaplab - a unified Sesame door to analysis tools, Proceedings, UK e-Science, All Hands Meeting 2003, Editors - Simon J Cox, p.509-513, ISBN - 1-904425-11-9, September 2003.

[17] 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.

[18] VOTable Format Definition http://www.ivoa.net/documents/VOTable/

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

Support of the workflow specifications reuse by ensuring its independence of the specific data collections and services

© Briukhov D.O., Vovchenko A.E., Kalinichenko L.A.

Institute of Informatics Problems (IPI RAN)

The paper is devoted to the problem of organization of the research process in the data-intensive sciences (DIS). It is focused on the problem of the workflow reuse. The paper presents an approach of embedding the subject mediators into the environment for collaborative research in DIS. This approach provides independence of problem solving methods and algorithms of the source data and services. It is shown that the independence of workflow from particular data collections and services constitutes a necessary requirement for the workflows re-use.