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).

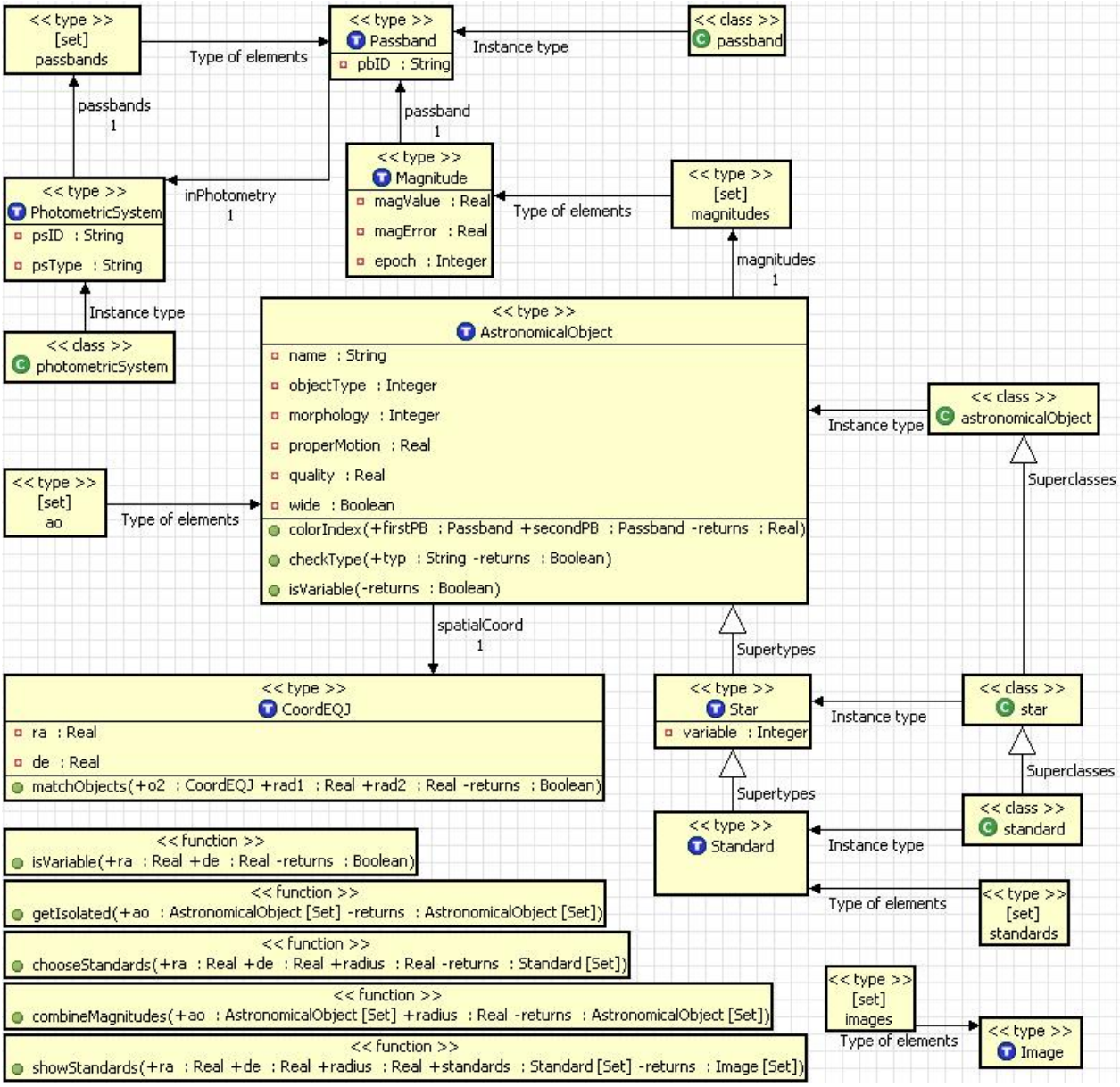

Рис. 1 Схема посредника для задачи определения вторичных стандартов

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.

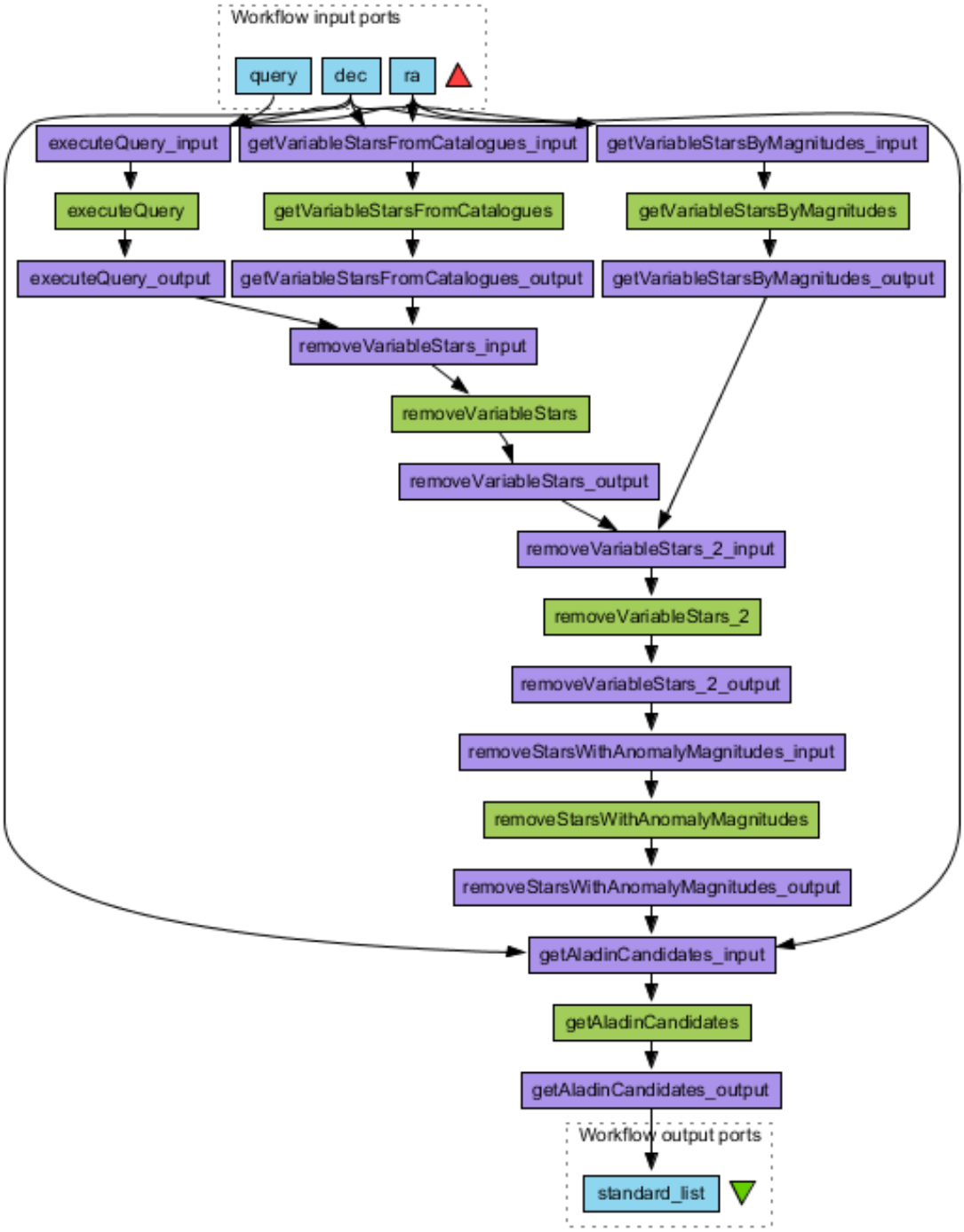

Рис. 2 Поток работ решения задачи вторичных стандартов в среде Taverna

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

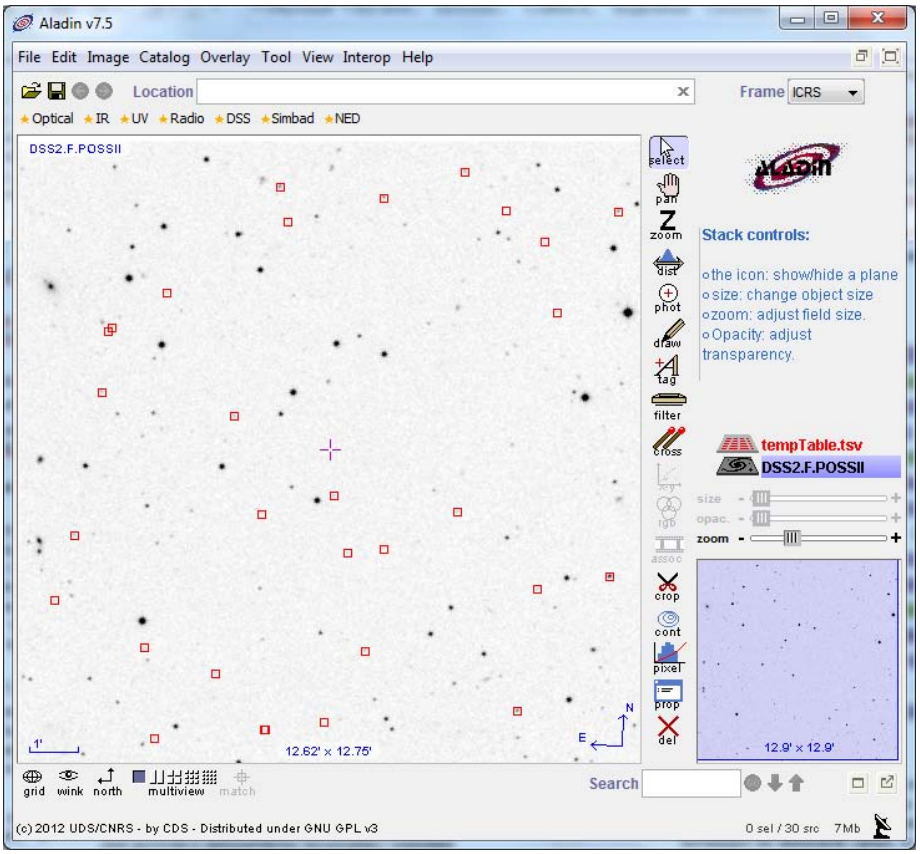

Рис. 3 Изображение найденных кандидатов в стандарты

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.

Метаданные о научных методах для обеспечения их повторного использования и воспроизводимости результатов

© Н. А. Скворцов, Д. О. Брюхов, Л. А. Калиниченко, Д. Ковалёв, С. А. Ступников
Институт проблем информатики РАН
Москва
nskv@ipi.ac.ru

Аннотация

В науках с интенсивным использованием данных предъявляются высокие требования к обработке больших объёмов данных набором научных методов для получения вторичной информации и новых знаний об исследуемых объектах. При этом важной оказывается доступность реализаций научных методов, применяемых в предметной области для организации обработки данных и решения задач. Обеспечение электронного хранения, повторного использования и воспроизводимости результатов экспериментов становятся неотъемлемыми атрибутами реализаций научных методов. В статье исследуется состав метаданных, которыми должны сопровождаться процессы, специфицирующие или реализующие научные методы, для обеспечения их повторного использования и воспроизводимости результатов. Компоненты процессов и данные сопоставляются с понятиями предметной области, сопровождаются информацией об их происхождении и качестве, системы тестов описывают разновидности ситуаций, в которых методы должны работать определённым образом. На примере открытой среды MyExperiment, организующей и предоставляющей доступ к коллекции научных потоков работ, показано, как расширение состава метаданных потоков работ позволяет организовать в коллекции семантический поиск релевантных решаемой задаче научных методов, проверить найденные реализации методов на интероперабельность, возможность повторного использования и обеспечить воспроизводимость результатов, полученных при их применении.

Работа выполнена при поддержке РФФИ (гранты 11-07-00402-а, 13-07-00579-а) и Президиума РАН (программа 16П, проект 4.2).

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

1 Введение

Получение колоссальных объёмов данных, подлежащих анализу научным сообществом, рождает качественное изменение в подходах к построению информационных систем для обработки данных и поддержки научных исследований. Науки с интенсивным использованием данных [1] призваны выявить полезные знания из объёма накопившихся ранее данных и потока появляющихся данных. Это требует постоянного автоматического применения широкого ассортимента известных методов, включая оценку существенных свойств и параметров объектов, проверку научных гипотез, выявление результатов, подтверждающих или опровергающих экспериментальные модели и так далее. Результаты применения научных методов сохраняются и становятся источником данных для работы других методов в данной области и сопряжённых проблемных областях.

Информационные системы в науках с интенсивным использованием данных комбинируют организацию информации в исследуемой области и организацию цифрового хранения и применения научных методов, используемых в данной предметной области. Научные методы могут представлять собой описание процессов обработки данных. Реализации методов разрабатываются в виде сервисов и потоков работ над доступными данными. Спецификации определяют, какие входные данные необходимы для работы методов, что и по каким алгоритмам они реализуют и какие результаты выдают. Потоки работ могут быть вложенными, то есть, вызывать друг друга в качестве подпроцессов.

Коллекции научных методов разрабатываются и занимают своё место в инструментах научного сообщества. В качестве примеров можно привести системы поддержки исследований в области астрономии. Проект VizieR [3] собирает всевозможные каталоги, организует их поиск и поиск в них, предоставляет набор сервисов, которые наиболее востребованы астрономическим сообществом. Однако до сих пор набор сервисов, реализующих какие-либо астрономические методы, достаточно ограничен. Этой информационной системой благодаря доступности данных пользуются практически все, кто работает с астрономическими данными. Среда виртуальной обсерватории Astrogrid [10], поддерживает удалённый доступ не только к данным, но и к сервисам различного назначения. Расширение Astrogrid средствами предметных посредников [12] позволило описывать в grid-среде спецификации предметных областей для формулирования и решения классов научных задач. Открытая коллекция научных потоков работ MyExperiment [6] объединяет тысячи пользователей и потоков работ и десятки проектов, в том числе в области астрономии, предоставляющих или использующих накопленные потоки работ.

Для того, чтобы подобные коллекции методов развивались и использовались научным сообществом, должна выйти на новый уровень вся инфраструктура поддержки научных исследований. Необходимо развитие и повсеместное использование сообществами общедоступных спецификаций предметных областей исследований и развитие семантических подходов решения задач с их использованием. Источники данных и реализации научных методов должны систематизироваться и связываться со спецификациями предметной области. Это позволяет упростить интеграцию информационных и методических ресурсов, автоматизировать многие шаги в обработке данных, которые до сих пор решались посредством ручных манипуляций всякий раз при решении новых задач. Реализации научных методов требуют разработки таким образом, чтобы упростить или даже автоматизировать их семантический поиск и использование в согласии со спецификациями предметной области. Данные и методы необходимо сопровождать информацией об их происхождении, точности, полноте. В цели разработки и реализации методов должны изначально закладываться возможность их повторного использования в данной и смежных областях, возможности воспроизведения результатов при одинаковых исходных данных. Создание инфраструктуры научных исследований, позволяющей использовать методы повторно, освободит исследователей от усилий, прилагаемых сегодня для интеграции неоднородных информационных ресурсов и реализации локально методов их обработки. Вместо этого само накопление методической базы, доступной, надёжной, согласованной со спецификациями предметной области и удобной в использовании, будет являться вкладом в развитие науке.

Данное исследование имеет целью разработку метаданных и методов работы с ними, которые должны сопровождать научные данные и реализации научных методов для достижения их повторного использования и воспроизводимости результатов научных экспериментов. В разделе 2 обсуждаются требования к доступным реализациям научных методов, исходным данным и получаемым результатам исследований в свете наук с интенсивным использованием данных. Раздел 3 посвящён связанным проектам и решениям. В разделе 4 более подробно описаны некоторые аспекты реализации проекта MyExperiment, выбранного для демонстрации возможностей инфраструктуры поддержки научных исследований с расширенным набором метаданных спецификаций научных потоков работ. Раздел 5 описывает собственно предлагаемый набор метаданных, сопровождающих доступные научные данные и реализации научных методов. В разделе 6 демонстрируется использование предложенных метаданных для организации поиска и повторного использования научных методов в инфраструктуре поддержки научных исследований.

2 Требования к реализации научных методов в среде поддержки научных исследований

Вначале необходимо представить требования, которые предъявляются к научным данным и методам для создания инфраструктуры поддержки научных исследований, позволяющей развивать спецификации предметных областей и коллекции научных данных и методов и использовать их реализации в исследованиях.

1. Под спецификацией предметной области, доступной и принимаемой сообществом исследователей, можно понимать набор связанных формальных онтологий предметной области исследования и смежных с ней областей. В соответствии с онтологиями могут создаваться концептуальные схемы предметной области, необходимые для организации информационных структур и спецификации методов, используемых в обработке данных.

Для развития семантических подходов к решению научных задач данные, информационные ресурсы и реализации научных методов необходимо связывать со спецификациями предметной области.

Агентами научного сообщества могут выступать как исследователи, так и информационные системы. Поэтому спецификации, описывающие методы и данные, должны обеспечивать понимание человеком и возможность машинной обработки. В этой связи необходимо использовать разработки, связанные с семантическим вебом [9].

2. Научные методы и данные должны быть отрыты и доступны для использования научным сообществом, работающим и решающим задачи в данной предметной области. Результаты работы методов также должны быть доступны для использования. Для этого они должны быть надлежащим образом специфицированы и опубликованы в общедоступных коллекциях. Коллекции собирают и систематизируют информацию и обеспечиваются средствами семантического поиска.

3. Важным принципом реализации научных методов является их независимость от источников данных. Подмена источников данных другими релевантными источниками надлежащего качества должна быть проста и не должна сказываться на работоспособности методов.

4. Для обеспечения повторного использования и данные, и методы необходимо сопровождать информацией об их происхождении. Она включает аутентификацию методов и данных, их источники, историю их развития и трансформации от создания до момента использования. С другой стороны, реализации методов должны сохранять информацию о происхождении обрабатываемых данных и обеспечивать дополнение этой информации в соответствии с манипуляциями, производимыми ими над данными.

5. Для оценки возможности повторного использования данных, методов и результатов расчётов или экспериментов необходима информация об их качестве: точности и полноте открытых данных, точности и полноте результатов, обеспечиваемых научными методами.

6. Обеспечение повторного использования также предполагает необходимость достаточно подробных спецификаций требований к их входным и выходным данным.

7. Обеспечение воспроизводимости результатов работы методов подразумевает под собой средства описания среды, необходимой для исполнения предоставляемых методов, спецификации поддерживаемых стандартов, а также наборы тестов, обеспечивающих проверку работы методов в различных ситуациях.

3 Связанные работы

Интересной разработкой с точки зрения накопления научных методов является среда разработки и сбора научных потоков работ MyExperiment [6]. Она организована как социальная сеть, позволяющая регистрировать исследователей, включать их в различные тематические группы, публиковать потоки работ, реализованные в различных сторонних системах, описывать эксперименты, связанные с вызовом потоков работ, составлять объекты исследования (фактически проекты), состоящие из потоков работ, документов, файлов данных, ссылок. Среда MyExperiment обеспечивает поиск потоков работ по метаданным, предоставляет их описание, позволяет их запускать. Интерфейсы среды соответствуют стандарту связанных открытых данных [11] и имеют соответствующие интерфейсы для этого. Тем временем, у данной среды есть ряд недостатков, препятствующих возможности повторного использования и воспроизведения результатов исполнения потоков работ.

То, что спецификации потоков работ публикуются в виде файлов, сгенерированных в форматах сторонних редакторов потоков работ, с одной стороны, позволяет использовать различные средства для их создания, с другой стороны, является причиной неоднородности и невозможности автоматизации использования опубликованных реализаций. В частности, спецификации потоков работ, созданные в наиболее используемом в данной среде внешнем редакторе Taverna [7], разбираются средой для выделения входных и выходных данных, визуализации структуры потоков работ, однако не имеет интерфейсов доступа к внутренней структуре потоков работ.

Данные для экспериментов и результаты, связанные с потоками работ, в MyExperiment также отданы на откуп внешним редакторам. В частности, Taverna поддерживает включение в спецификацию потока работ тестового примера для исполнения. Для подтверждения воспроизводимости результатов этого недостаточно, так как невозможна спецификация различных случаев и альтернативных путей прохождения потока работ.

В среде MyExperiment нет требования независимости методов от источников данных или возможности подмены источников, и в коллекции есть множество потоков работ, которые по своей сути являются не реализациями методов, а сервисами, предоставляющими данные из специфических источников данных по некоторым входным параметрам.

Хотя MyExperiment декларирует расширяемость онтологии, на которой построена схема информационной системы, на деле связи спецификаций потоков работ с какими-либо описаниями предметной области исследования сделать посредством существующих интерфейсов невозможно. В среде поддерживаются только вербальные пояснения к потокам работ и теги, и обеспечивается возможность поиска по ним.

В Taverna поддерживаются спецификации происхождения данных. Однако предназначены метаданные о происхождении только для записи пути прохождения данных внутри исполненного потока работ. Для достоверной проверки возможности повторного использования данных этого явно недостаточно, так как невозможно отследить историю их получения и преобразования от момента создания. К тому же доступа через интерфейсы MyExperiment к имеющимся данным о пути преобразования данных в потоке работ нет.

Проект wf4ever [4] предоставляет набор средств для поддержки повторного использования, проверки применимости, воспроизводимости и других свойств потоков работ. Среди описаний в проекте возможно специфицировать происхождение, внутреннюю структуру потоков работ, возможности доступа, жизненный цикл, развитие, многоверсионность и другие аспекты. Потоки работ могут проверяться на полноту, непротиворечивость, доступность и совместимость источников данных. Для этого предоставляются необходимые структуры данных и интерфейсы пользователя. В данном проекте в качестве экспериментальной базы взята коллекция потоков работ MyExperiment. Спецификации предметов исследования и потоков работ можно импортировать из MyExperiment, дополнить спецификациями, предоставляемыми проектом, и использовать набор сервисов для поддержки жизненного цикла потоков работ. Проект не предполагает больших продвижений в сторону семантических подходов к обеспечению доступа к потокам работ, а направлен больше на анализ самих потоков. В частности, одной из целей экспериментов ставится анализ того, почему многие из потоков работ в среде MyExperiment на сегодняшний момент попросту не запускаются.

4 Среда поддержки коллекции научных потоков работ MyExperiment

На примере среды разработки и публикации научных потоков работ MyExperiment мы будем показывать, какие метаданные необходимо добавлять к спецификациям потоков работ для обеспечения их повторного использования и воспроизводимости результатов. Поэтому более подробно остановимся на реализации сред MyExperiment

Для хранения метаинформации о потоках работ в среде MyExperiment используется база данных, схема которой специфицирована набором модулей онтологии. В этих модулях определены средства описания внутренней структуры накапливаемых потоков работ, спецификации пользователей, групп, аннотаций и других необходимых метаобъектов. Рассмотрим часть из них, представляющую интерес для данного исследования.

Для хранения метаобъектов о различных видах компонентов потоков работ создано базовое понятие WorkflowComponent. Его подпонятие NodeComponent описывает узлы потоков работ. Разновидности узлов представлены понятиями: Source – узлы-источники, приносящий в поток работ данные на обработку, Sink – узлы окончания потока работ, в которые приходят данные результатов исполнения потока работ., и Processor – узлы, исполняющие сервисы обработки данных. В свою очередь, типы исполнительных узлов описываются подпонятиями. В частности, WSDLProcessor соответствует вызову веб-сервиса. DataflowProcessor специфицирует вложенный поток работ, также состоящий из компонентов. Данные, Входы, выходы и соединения каждого узла в потоке работ описываются понятиями Input, Output и Link соответственно и объединяются базовым понятием IOComponent.

Объект исследования в MyExperiment представляет собой контейнер, содержащий файлы (например, данные, документы), внешние ссылки и потоки работ. Для хранения потоков работ как целостного объекта служат понятие AbstractWorkflow и его подпонятия Workflow и WorkflowVersion. Аналогично спецификациям файлов соответствуют понятия AbstractFile с подпонятиями File и FileVersion. Такая организация позволяет создавать многоверсионные объекты.

Понятия файлов и потоков работ объявляются имеющими суперпонятия Annotatable. С помощью этого понятия с ними могут быть связаны несколько видов аннотаций, среди которых комментарии, цитирования, теги и другие. Теги используются в качестве описания потоков работ и файлов для поиска в коллекции MyExperiment.

Сами метаобъекты, описывающие потоки работ, хранятся в реляционной базе, но реализована генерация их представления в модели RDF как экземпляров онтологии MyExperiment. Каждый метаобъект имеет в системе свой уникальный идентификатор URI. Например, идентификатор конкретного потока работ выглядит так: http://www.myexperiment.org/workflows/3514/.

Для разработчиков приложений над MyExperiment доступны несколько интерфейсов. К метаинформации MyExperiment можно задавать http-запросы через REST-интерфейс. Java-интерфейс MyJPI представляет собой REST-интерфейс, обёрнутый в классы языка Java. Наконец, реализован интерфейс точки доступа SPARQL, позволяющий задавать запросы к метаинформации MyExperiment и получать RDF-данные в соответствии со схемой, заданной онтологией, в нескольких форматах с учётом или без учёта автоматического вывода по правилам RDF Schema.

Однако все упомянутые интерфейсы имеют ограничение: в них не реализован доступ к внутренней структуре потоков работ, несмотря на то, что она определяется онтологией как компоненты потоков работ. Посредством программных интерфейсов можно получить ссылку на поток работ как файл Taverna. Этот файл подлежит разбору уже средствами Taverna для получения данных о внутренней структуре потоков работ. Это означает, что в рамках запроса на получить внутреннюю структуру потока работ не удастся.

В составе объектов исследования, помимо файлов (документации, данных), ссылок, потоков работ и аннотаций, поддерживаемых в MyExperiment, для обеспечения требований, изложенных в разделе 2, должны содержать также исчерпывающие наборы тестов, учитывающие

различные ситуации, и соответствующие данные результатов тестов при разных входных условия.

Таким образом, для создания среды исследований, обеспечивающей семантический поиск методов, повторное использование и воспроизводимость, в MyExperiment требуется расширение интерфейсов доступа к структуре потоков работ и поддержка систем тестов с результатами. В целом, это возможно, так как MyExperiment является проектом с окрытым кодом. Однако на данном этапе исследование проводилось с использованием оригинального сервера MyExperiment, соответственно, средства со стороны MyExperiment не менялись.

5 Расширение состава метаданных, сопровождающих публикуемые данные и научные методы

Для поиска объектов исследования, релевантных решаемой задаче, в MyExperiment предназначены только их текстовые описания и аннотации тегами. Причём связаны они, могут быть только с потоками работ в целом или файлами, исходя из их суперпонятия Taggable. Для коллекции методов и потоков работ, обеспечивающей их повторное использование, этого, безусловно, недостаточно.

Мы производим расширение состава хранимых метаданных об объектах исследования, потоках работ и их компонентах, для реализации семантических подходов работы с методами предметной области. Спецификации расширенного состава метаданных оформляются в виде набора онтологий разного назначения. Описанные онтологические модули находятся в открытом доступе по адресу: http://ontology.ipi.ac.ru/ontologies/astront, – и могут использоваться для накопления метаданных в соответствии с их определениями. Для хранения метаданных, связанных с конкретными метаобъектами MyExperiment, используется отдельная база экземпляров RDF.

Для реализации семантических подходов к поиску потоков работ, релевантных решаемой задаче, их повторному использованию и обеспечению воспроизводимости, в первую очередь, необходимо развивать спецификации предметной области, в которой собирается коллекция методов. Поиск потоков работ, отвечающих требованиям задачи, необходимо связывать с онтологией предметной области, которой принадлежит коллекция и в которой решается задача. Для этого метаобъекты, описывающие потоки работ, объявляются экземплярами классов понятий онтологии предметной области. Отнесение метаобъекта к классу понятия в терминах онтологий реализуется посредством отношения rdf:type. Для более сложных описаний в терминах онтологии метаобъекты могут становиться экземплярами неименованных классов, определённых как подпонятия понятий онтологии, но без введения новых понятий и свойств в онтологию.

Мы рассматриваем предметную область звёздной астрономии, включающую понятия одиночных звёзд, кратных систем звёзд. С ними связаны модули с описанием понятий астрометрии, фотометрии, астрофизики как понятий смежных областей. Эти модули используются в большинстве задач в области астрономии вне зависимости от того, какие задачи они решают.

В частности, в модуле астрометрии определены следующие понятия:

- Coordinate
- CoordinateSystem
- EquatorialCoordinateSystem
- CoordinateSystemComponent
- Epoch
- RightAscension
- Declination
- и другие.

Понятия имеют иерархию, описание структуры с помощью связей и ограничений.

В онтологию предметной области включены также более специфические модули, определяющие знания о парах и компонентах кратных звёзд, параметрах орбит двойных звёзд, параметрах кривой светимости затменных звёзд и других. Такие модули используются в более узких классах задач, в частности, связанных с определёнными видами астрономических объектов.

В качестве примера отнесения данных или компонентов потоков работ к понятиям онтологии предметной области, метаобъект с данными о координате прямого восхождения (RA_J2000) астрономического объекта может быть связан с понятием онтологии RightAscension, но для более точного описания такой метаобъект должен стать экземпляром выражения (подпонятия) в терминах онтологии, ограничивающего класс множеством экземпляров *x* таких, что x принадлежит Right Ascension, и существует координата *y*, система координат у которой экваториальная, и у которой есть компоненты: *x* и эпоха, равная J2000. Выбор простого или более точного стиля описания метаданных в дальнейшем влияет на качество поиска метаобъектов в терминах онтологии.

Наряду с модулями онтологии предметной области в нашем подходе спецификации метаданных пополняются также специализированными онтологиями, описывающими требования к происхождению данных, их качеству и среде исполнения.

В качестве онтологии происхождения данных используется в соответствии с рекомендацией W3C онтология PROV-O [2]. В её основе лежат понятия агента (Agent), деятельности (Activity) и сущности (Entity). Агентами могут быть человек (Person),

организация (Organization) или программа (SoftwareAgent). Вариации отношений их экземпляров друг с другом описывают различные события и ситуации, которые необходимо фиксировать при преобразовании, перемещении, изменении статуса данных. Например, метаданные об исходных данных, которые использовались процессом, выражается отношением used, связывающего агента и деятельность; информация об инструменте, который был использован для генерации результата, выражается отношением wasAttributedTo, связывающего сущность и программу и так далее. Посредством такой онтологии можно задавать метаданные об авторстве и принадлежности данных и методов, прослеживать историю преобразования данных от первоначального источника до текущего состояния, сопровождать реальные данные и методы другой подобной информацией.

Приведём пример спецификации происхождения данных для потока работ wf3514, обращающегося к внешнему сервису resolve_coordinates (Sesame Name Resolver) для локализации астрономического объекта на небе по его имени. Результирующие данные потока resolve_coordinates_outputTable могут содержать информацию в виде триплетов об инструменте, которым созданы данные и о потоке работ:

```
wf3514:resolve_coordinates
    rdf:type prov:SoftwareAgent .
wf3514:resolve_coordinates_outputTable
    rdf:type prov:Entity;
    prov:wasAttributedTo
        wf3514:resolve_coordinates;
    prov:wasGeneratedBy wf3514:wf3514 .
```

Ещё одна часть спецификации необходимых метаданных, онтология качества данных DQ [5], содержит набор факторов качества данных, определяемых измерениями в многомерном пространстве значений и метриками качества в этих измерениях. В качестве примера взяты измерения полноты данных (Completeness), объёма данных (Data Volume), возраста данных (Timeliness), точности (Accuracy), целостности (Consistency), меры доверия (Confidence). Состав измерений и метрики для их реализации сильно зависят от предметной области исследования. С одним объектом может одновременно быть связано несколько значений качества в разных измерениях. Экземпляры понятий данной онтологии связываются с потоками работ и файлами в целом, любыми компонентами потоков работ, сервисами и их параметрами, а также с самими данными. Метрики оценки качества также могут различными, но они согласовываются и специфицируются сообществом, работающим в предметной области.

Спецификации сред воспроизведения также могут требовать определения некоторой структуры метаданных. Однако, данные, необходимые для обеспечения воспроизводимости экспериментов, в многом выразимы средствами онтологии происхождения данных.

Также в среде MyExperiment требуется разработка поддержки систем тестов. До сих пор они описываются только некоторыми исследователями и неформально, в поле описания потока работ, либо в файлах, включённых в коллекцию объекта исследования. После реализации такой поддержки входные и выходные данные тестов, должны связываться

Для соответствия разработанным требованиям к публикации научных методов необходимо обеспечение определённых метаобъектов MyExperiment метаданными в терминах упомянутых онтологий.

Метаданными в терминах онтологии предметной области должны сопровождаться:

- файлы, потоки работ как целостные объекты;
- входные узлы в качестве предусловий;
- выходные узлы в качестве спецификаций их постусловий;
- узлы обработки данных;
- их входы и выходы.

Таким образом, производится описание семантики компонентов потоков работ в онтологии, на основе которого появится возможность поиска потоков работ, релевантных задачам, по понятиям, соответствующим потокам в целом, по соответствию семантики входных и выходных узлов, по семантике узлов обработки, по семантике блоков и потоков данных внутри потоков работ. Помимо поиска появляется возможность верификации потоков работ и их использования.

Метаданными в терминах онтологии происхождения сопровождаются:

- сами потоки работ как описания научных методов, требующих прояснения происхождения;
- обрабатывающие компоненты потоков работ как определённые научные сервисы;
- данные, направляемые на обработку в потоке работ, находящиеся в процессе обработки и результирующие.

Любые данные, входящие в объект исследования в виде файлов или участвующие в потоках работ, должны быть соотнесены с онтологиями предметной области, происхождения, качества данных.

Некоторые аспекты качества данных могут быть связаны с методами и потоками работ в целом как спецификациями качества, ожидаемого от работы методов.

Тесты и их результаты снабжаются связями с онтологией предметной областью, причём особенности различных ситуаций, представляемых разными тестами, желательно отражать в

ограничениях понятий. Результаты тестов должны иметь метаданные происхождения, связанные с историей выполнения тестов в потоках работ.

6 Применение метаданных для обеспечения повторного использования и воспроизводимости результатов работы научных методов

Онтологии предметной области исследования, происхождения данных, качества данных, сред исполнения фактически определяют разные ракурсы взгляда на описываемые объекты исследования и научные методы. Метаданные в терминах определённых онтологий – не зависимые друг от друга проекции на объект исследованияв контектсе знаний данной онтлогии. Запросы в терминах каждой из этих онтологий, могут выдать потоки работ или их компоненты, соответствующие определённым требованиям с точки зрения конкретной онтологии.

Для хранения метаданных используется база RDF-триплетов на основе Jena. В ней хранятся экземпляры в соответствии со структурой, определённой описанными выше онтологиями. Для работы с базой экземпляров используется язык запросов SPARQL.

При решении научных задач и поиске релевантных задаче реализаций научных методов возникнет необходимость предъявления требований одновременно с нескольких ракурсов. Таким образом, понадобится обрабатывать запросы, включающие конъюнктивно требования одновременно в терминах нескольких онтологий.

Пример запроса.

```
prefix rdf:
<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix mecomp:
<http://rdf.myexperiment.org/ontologies/components/>
prefix astrobjects:
<http://ontology.ipi.ac.ru/ontologies/astrobjects.owl>
prefix prov:
<http://www.w3c.org/ns/prov#>
SELECT ?workflow WHERE
{ ?output rdf:type astrobjects:AstrObject .
  ?output prov:wasGeneratedBy ?workflow .
  ?output prov:wasAttributedTo :resolve_coordinates.
  SERVICE <http://rdf.myexperiment.org/sparql>
  { ?output mecomp:belongs-to-workflow ?workflow .
    ?output rdf:type mecomp:Sink }
}
```

Такой запрос к базе RDF-экземпляров выясняет, какие потоки работ из коллекции MyExperiment возвращают астрономические объекты, обращаясь за ними в сервис resolve_coordinates (с точки зрения онтологии происхождения данных). Соответственно, он включает в себя требования к выборке из точки доступа MyExperiment метаобъектов класса Workflow, к которым относятся метаобъекты класса Sink. В языке запросов SPARQL для обращения к распределённым точкам доступа используются средства федеративных запросов с помощью конструкции SERVICE. и прямого указания адреса точки доступа MyExperiment. Остальные требования относятся к тем же RDF-ресурсам, но опрашивается база RDF-экземпляров с метаданными. Одно из них относится к метаданным в терминах онтологии астрономии, а именно, принадлежность выходных данных потока работ понятию AstrObject,. А другое – к метаданным в терминах онтологии происхождения данных, а именно, какой инструмент используется для генерации данных. Таким образом, один запрос использует термины MyExperiment, термины онтологии предметной области и термины происхождения данных, а результатом запроса являются найденные в коллекции научных методов потоки работ, релевантные сформулированным в запросе требованиям.

Подобное использование метаданных позволяет решать многие задачи, связанные с семантическим подходом к обеспечению интероперабельности научных методов, их повторным использованием и обеспечением.

На основе метаданных о связи с предметной областью можно решать задачи поиска релевантных методов:

- по понятиям, связанным с потоками работ в целом;
- по соответствию требованиям задачи понятий, связанных с входными и выходными данными потоков работ, то есть, по спецификации в терминах онтологии того, что мы имеем и того, какие данные мы имеем, и того, что хотим получить в результате работы метода;
- по присутствию в потоке работ компонентов-стадий, которые необходимы для решения задач;
- по другим возможным критериям, формулируемым с использованием понятий предметной области.

Возможно производить семантический контроль используемых методов и принятых решений:

- проверку семантики данных между всеми компонентами потока работ;
- проверку корректности использования подпроцессов по их входным и выходным параметрам;
- соответствие семантики входного компонента семантике входных данных, либо выходных данных выходным компонентам;
- соответствие семантики данных, проходящих из выхода одного компонента на вход другой, по принципу спецификаций пред- и постусловий:

постусловие выхода предыдущего компонента должно быть строже предусловия входа последующего компонента.

Видно, что обеспечение семантической интероперабельности за счёт соотнесения задач, данных и методов со знаниями предметной области является основой для обеспечения повторного использования научных методов.

Обеспечение качества данных, достоверности, полноты и других аспектов, связанных с надёжностью данных и методов, реализуется с помощью использования онтологиями качества данных и их происхождения.

Возможности метаданных происхождения данных также сложно переоценить. С их помощью осуществляется:

- контроль реальных источников данных и их качества в соответствии с требованиями задачи;
- контроль за соответствием требованиям решения задачи используемых открытых реализаций научных методов
- контроль прохождения тестов по определённому пути в потоках работ и соответствия качества получаемых данных требованиям задачи
- проверка требований воспроизводимых экспериментов к исполняемой среде.

Таким образом, воспроизводимости результатов способствует ведение метаданных происхождения для каждой манипуляции, производимой при прохождении экспериментов. При воспроизведении результатов возможно отследить обратную цепочку манипуляций и повторить её.

Спецификации требований к исполняемой среде, необходимой для проведения эксперимента, формулируются в терминах происхождения данных.

7 Заключение

В статье проанализированы требования к средам поддержки научных исследований для обеспечения повторного использования научных методов и воспроизводимости результатов их работы. Предложен набор метаданных, которые должны сопровождать данные и методы с этой целью. Метаданные определяются в терминах онтологий и включают привязку описаний научных методов и потоков работ к знаниям предметной области и также снабжение информацией о происхождении и качестве данных. Показан путь использования этих метаданных.

Литература

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

[2] The PROV Ontology. W3C Recommendation. – W3C, 2013. – URL: http://www.w3.org/TR/prov-o/.

[3] VizieR. – URL: http://vizier.u-strasbg.fr/cgi-bin/VizieR

[4] Wf4Ever project. – URL: http://www.wf4ever-project.org/

[5] S. Geisler, S. Weber, Ch. Quix. Ontology-based data quality framework for data stream applications. // Proc. of the 16th International Conference on Information Quality (ICIQ-11). – 2011.

[6] Goble C. A., De Roure D. C. myExperiment: social networking for workflow-using e-scientists // Proceedings of the 2nd workshop on Workflows in support of large-scale science. – ACM, 2007. – С. 1-2.

[7] D. Hull, K. Wolstencroft, R. Stevens, C.A. Goble, M.R. Pocock, P. Li, T. Oinn. Taverna: A tool for building and running workflows of services, Nucleic Acids Research, 34 (Web-Server-Issue), 2006, pp. 729–732.

[8] L. Moreau. Provenance-Based Reproducibility in the Semantic Web. // Web Semantics: Science Services and Agents on the World Wide Web. – 9, (2). – 2011. – P. 202-221.

[9] Shadbolt N., Hall W., Berners-Lee T. The semantic web revisited //Intelligent Systems, IEEE. – 2006. – Т. 21. – №. 3. – С. 96-101.

[10] Walton N. A. et al. AstroGrid: A place for your science //Astronomy & Geophysics. – 2006. – Т. 47. – №. 3. – С. 3.22-3.24.

[11] Yu L. Linked open data //A Developer's Guide to the Semantic Web. – Springer Berlin Heidelberg, 2011. – С. 409-466.

[12] А. Е. Вовченко, Л. А. Калиниченко, С. А. Ступников Семантический грид, основанный на концепции предметных посредников. // Труды четвертой международная конференция "Распределённые вычисления и Грид-технологии в науке и образовании" Grid2010, Дубна, ОИЯИ, 2010. – с. 309-318.

Metadata of scientific methods for achievement of their reuse and reproducibility of their results

N. A. Skvortsov, D. O. Briukhov, L. A. Kalinichenko, D. Kovalev, S. A. Stupnikov

In data-intensive sciences there are high requirements imposed on big volume data processing with a set of scientific methods to achieve secondary information and new knowledge about investigated objects. So accessibility of scientific method implementations being applied at the domain for data processing and problem solving organizing comes out to be important. Digital reservation, reuse and reproducibility of experiment results become inherent attributes of scientific methods. The paper investigates metadata structure to be attached to processes

specifying or implementing scientific methods for their reuse and result reproducibility. Process components and data are referred to the domain concepts and are supplied with information about their provenance and quality, test collections describe kinds of situations in which methods must behave in a definite way. On the case of the open MyExperiment environment organizing and providing access to the collection of scientific workflows we demonstrate how the extension of metadata allows to organize at the collection semantic search for methods relevant to a problem, to verify interoperability, reusability and reproducibility of method implementations.

.

Core semantic model for generic research activity[€]

© Vasily Bunakov
Scientific Computing Department, Science and Technology Facilities Council,
Harwell OX11 0QX, United Kingdom
vasily.bunakov@stfc.ac.uk

Abstract

A simple research activity model is suggested that is agnostic to research domain and allows independent curation of the research information lifecycle by a variety of its stakeholders with a potential to further link individual activities into meaningful research provenance or research value chains. We consider the drivers for conceiving the model, its main aspects, an RDF manifestation of it, a particular business case for its application, and discuss its potential for future applications.

1 Introduction

Different stages of the research lifecycle in natural sciences as well as in social and economic research produce multiple data artefacts under control of different data management solutions and software platforms. (We use the term "data" here and there in a broad sense: not necessarily numeric data resulting from measurements but research proposals, software components, configuration files, electronic publications, etc.) Data curators working in a particular research domain tend to develop a specific metadata model that aims to cover the entire research lifecycle from the research inception to the research outputs dissemination. Such a metadata model quite often serves as a foundation for the design of the actual information systems and services. The example of a comprehensive metadata model for the research performed at large facilities like synchrotrons, powerful lasers or neutron sources is the Core Scientific MetaData model [5]; the example in social research is DDI-Lifecycle [7].

Proceedings of the 15th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" — RCDL-2013, Yaroslavl, Russia, October 14-18 2013.

[€] This work is related to the ENGAGE project www.engage-project.eu and the projects of PaNdata collaboration www.pan-data.eu supported by the EU 7th Framework Programme for Research and Technological Development. The author would like to thank his colleagues in ENGAGE and PaNdata for their input for this paper although the views expressed are the views of the author and not necessarily of the projects.

Substantial effort of renown information experts has been spent in order to extend some established metadata models with new semantic features; the example in social research will be DDI semantic modelling ([8], [9]). The richness and the expressivity of metadata model that has evolved through decades can be considered a limitation that makes it harder to agree on what should constitute the "true" semantic representation, or what format of it should be a "canonical" one. Also the attempts to transform the entire domain-specific metadata model into semantic representation, and then offer it for common adoption and data linkage may contradict the social nature of Linked Data as its curation can be reasonably considered an incremental and opportunistic effort of multiple parties (as brilliantly illustrated by [1]).

This is not to say that semantic modelling of the entire research domain is not sensible or do not have a potential for implementation. Collaborative projects of a multinational scale such as PaNdata-ODI ([2], also see under [16]) consider semantic representation of the popular domain-specific metadata model [5] with the purpose of system integration. The motive for this consideration is that, despite the actual information systems in different research centres may be based on the implementations of the same generic metadata model and even on the same software platform for data catalogue [14], the practices of the catalogue configuration, the interpretation and the use of the model elements, and hence the actual semantics of these elements may vary dramatically. A common semantic layer, probably in the form of ontology, is considered then a viable architecture solution that should allow retaining the existing local practices of data cataloguing and at the same time, should give the IT teams an ability to meaningfully integrate distributed data and services.

That semantic layer, however, will require an inclusion into a certain best practices framework to sustain it through time [4], otherwise divergent business needs and business practices of the collaboration participants can make a thoroughly designed semantic model obsolete the next day after its implementation in a real IT solution. Keeping a comprehensive semantic model actual can be quite an expensive endeavour with substantial overheads on continuous business analysis and communication with multiple parties.

Another concern about the attempts of semantic representation of comprehensive metadata models is a tendency for them to reflect the information needs of

only a few types of the research lifecycle stakeholders: this is commonly Researchers and Data Archivists. The information needs of other stakeholders from Funding, Industry, or Education are often under-represented. To resolve this issue, one can take two approaches:

A) As a responsible information curator, conduct thorough business analysis of the research lifecycle stakeholders' types and their information needs then incorporate the knowledge acquired into a comprehensive model that, in order to be effective, should be validated by the stakeholders themselves (then, ideally, permanently amended).
B) Give different stakeholders a reasonable modeling means to express their role in the research lifecycle so that each of them becomes an information curator who cares about the quality and the actuality of her contribution into the shared pool of information.

The latter approach seems more adequate in the present situation when the advance of Linked Data principles allows various stakeholders to meaningfully model their part of information universe, also re-use the results of similar modeling effort made elsewhere.

We suggest a small but quite universal "core" model in the spirit of Linked Data principles [1] with low barriers for its adoption and use for semantic annotation of the research activity in different local information contexts, with their further inclusion into a global information context. We think that such a model should not focus on data but on common patterns of research activity observed in different research domains (for which we give examples further in this paper); various data then can be considered artefacts or "footprint" of different types of research activity.

2 Research activity model

2.1 Types and common patterns of research activity

Research lifecycles analyzed and structured by digital curators in the respective research domains can be a good source for discovering granular research activities and their interrelations. In this work, we consider two lifecycles: in facilities science[1] and in social research; they are most relevant to the projects which contributed to the development of our model ([11], [16]) and their respective research domains stay quite far apart so may help us with testing our model universality.

Lifecycle in facilities science that underpins CSMD model [5] includes the submission of a research proposal to the facility user office in order to get the facility resource for research (e.g. beam time on synchrotron); the further approval of the proposal by the facility's user office; experiment scheduling; conduct of the actual experiment with data collection; data storage; data analysis; and eventually publishing research results with record keeping for them. Beyond this lifecycle that is supported by facility itself, there is research funding activity, or research policy making, or the researchers' social communication that all can be considered elements of a larger "research value chain".

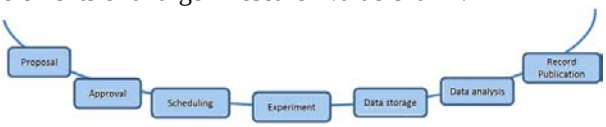

Figure 1. Research lifecycle in facilities science (as captured by CSMD model).

The lifecycle of social research that underpins DDI-Lifecycle model [7] includes the formulation of the study concept, further data collection, its processing, archiving, distribution, discovery, analysis, and repurposing. Funding, or policy making, or social communication, despite there are some placeholders for references to these types of activity – are again beyond the immediate scope of DDI.

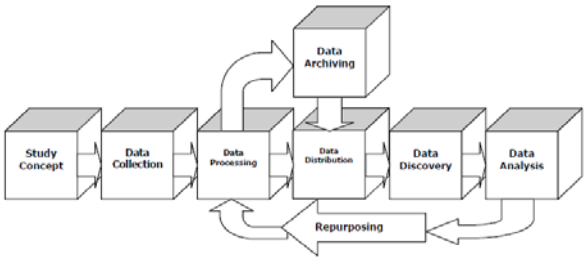

Figure 2. Research lifecycle in social science (as captured by DDI-L model).

Each activity yields certain outputs, e.g. in facilities science, the research proposal preparation results in the investigation (experiment) description, data analysis yields derived data etc. Previous activity may provide an input for other activity or give it a context, e.g. it is quite common for researchers to refer to the previous investigations (experiments) when they apply for a new investigation to be conducted at the same facility.

Despite there are similarities between the two aforementioned lifecycles and between the roles of stakeholders involved in them, there are differences, too. Even more differences come up if we consider context or scope of each research activity, or means for their description that are present in each model. As an example, in facilities science, the scope of experiment can be understood by considering what samples or chemical substances have been under investigation; in social research, it can be meaningful parameters describing the human audience which the study has been aimed upon. Not these details that may be different but the very presence of Context and Scope, as well as the Inputs and Outputs for the research activity, or Actors who perform it, or Effects of the research do represent a common pattern – very generic but universal

[1] For the sake of clarity, we use the term "facilities science" for the research performed on large-scale scientific instruments (synchrotrons, powerful lasers and alike) by visitor teams or individual researchers who obtain, via the application process, access to the common facility resource in order to conduct their experiments or observations, and to collect the resulting data.

across research fields.

These patterns are common not only across different research domains for the similar types of research activity (when we draw parallels e.g. between facility science Experiment and social research Study); this is also the case for different types of research activity within the same lifecycle, e.g. funding or data analysis or record publication have their Inputs and Outputs, their Actors, Effects, Context (Conditions) and Scope.

These basic patterns contribute to a reasonable model that should not be too burdensome for the respective stakeholders (or information specialists working for them) to apply, yet is expressive enough to promote the principles and best practices of Linked Data in various research domains. We consider a potential for such an application below in the section devoted to a particular business case; in the meanwhile, we are going to formally introduce the major aspects of a generic research activity, and suggest a practical RDF-based manifestation for them.

2.2 Generic research activity (research activity "cell")

We deem important the following aspects of a generic research activity:

Aspect	Description	Examples	
		Research per se	Research data analysis
Input	Something that is taken in or operated on by Activity	Previous research	Raw data
Output	Something that is intentionally produced by Activity	Raw data	Derived (analyzed) data
Scope	Something that Activity is aimed at or deals with	Sample properties	One or more experiments
Condition	Something that affects or supports Activity, or gives it a specific context	Scientific instrument	IT environment
Actor	Something or somebody who participates in Activity	Investigator	Data analyst
Effect	Something that is a consequence of Activity	Environment pollution	New software module

Schematically, the granular research activity can be represented by the following diagram:

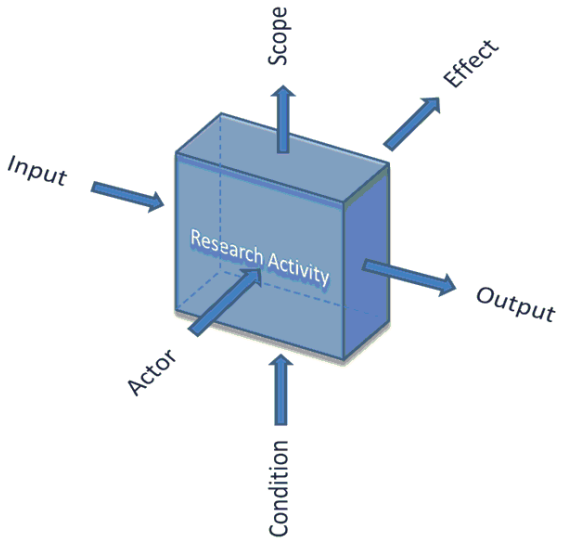

Figure 3. Research activity "cell".

Research activities can be combined as "cells" in chains where Output of one can be an Input to another but in fact, the model allows other sorts of links between activities. As an example, a piece of regulation such as data management policy can be an Output of one activity (policy making), and a Condition that affects another activity (research per se); a new software module that is a side Effect of a certain activity (data analysis) can be a non-human Actor that participates in other activity (e.g. automated indexing of experimental data). This shows that activity aspects in fact do not have "types": a modeler can use and combine them as dictated by the semantics of the respective subject area.

This view is inspired, to some extent, by SADT activity model [17] with its idea of combining activities into the hierarchy or a grid but is quite different by introducing some other activity aspects and not imposing their typization. Also SADT promotes a top-down approach to structured analysis and systems design when we suggest a bottom-up approach that allows combining the granular activities in more complex information structures.

Compared to other project-driven attempts to model research activity ([10], [15]) our model is going to be simpler, more universal, and deliberately aimed at semantic modeling of a granular activity rather than of the entire research lifecycle thus providing a "building block" for a more sophisticated information modeling as and when required.

2.3 RDF manifestation of activity model

The outlined model may imply different manifestations; we feel that one expressed in RDFS Plus (RDF Schema with a few OWL terms) has a good potential for adoption by information curators and implementation in real IT solutions. This paper Appendix suggests the

RDFS Plus manifestation of the activity model that can be extended by domain specific entities and properties. As an example, an information modeler in facilities science might want to extend the model as follows:

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix am: <http://example.org/stuff/ActivityModel#>.
@prefix rm: <http://example.org/stuff/ResearchModel#>.
# For Activities
    rm:Research rdfs:subClassOf am:Activity .
    rm:Experiment rdfs:subClassOf rm:Research .
# For Conditions
    rm:Condition rdfs:subClassOf am:Condition .
    rm:Regulation rdfs:subClassOf rm:Condition .
    rm:DataManagementPolicy rdfs:subClassOf rm:Regulation .
# For Output
    rm:Output rdfs:subClassOf am:Output .
    rm:Publication rdfs:subClassOf rm:Output .
    rm:Dataset rdfs:subClassOf rm:Output .
# For Scope
    rm:Scope rdfs:subClassOf am:Scope .
    rm:ExperimentalTechnique rdfs:subClassOf rm:Scope .
    rm:SubjectCoverage rdfs:subClassOf rm:Scope .
# For properties
    rm:activity_location rdfs:subPropertyOf am:hasScope .
    rm:activity_subject rdfs:subPropertyOf am:hasScope .
```

The user of the information system where the RDF data prepared according to our model is published can then use reasonable SPARQL requests to inquire for different aspects of research activities, e.g. trying to realize first how much research output, and how much of each type is out there:

```
SELECT ?output_type (COUNT(?output) as ?total)
WHERE {?output_type rdfs:subClassOf am:Output .
  ?output a ?output_type .
    }
GROUP BY ?output_type
```

or try to discover the chains of interrelated activities:

```
SELECT ?previous_activity ?current_activity
WHERE {?previous_activity am:hasOutput ?output .
     ?output am:inputFor ?current_activity .}
```

User may be familiar with just our activity model knowing very little about a certain research domain at start, then accumulating more and more knowledge through sensible incremental requests. In case the information modeler, in addition to our basic activity model, has followed good practices of data curation so that e.g. instances of Scope or Condition subclasses are not literals but dereferenceable URIs, the User will have even more opportunities of getting familiarized with the semantics of a particular research domain. When we tell of "User" we of course mean the software agents, too, as the prospect of employing them is a strong incentive for any semantic modeling.

2.4 Business case for semantic categorization and annotation of existing metadata

As we mentioned, it may not be easy to give birth to the semantic representation of a comprehensive metadata model because of its richness and complexity, and because of substantial overheads for communication among information curators who apply the model in different contexts. Another observation is that detailed metadata records may in fact represent different activities performed by different stakeholders of the research information lifecycle – while the records that in fact circulate in the information management solutions are focused on particular types of stakeholders only and support their specific roles in the first place. A certain stakeholder, e.g. Data Librarian or Data Archivist may claim that Her information management solution is focused on *data* in pursuit of some common interest when, in fact, the information management solution primarily supports this particular stakeholder specific *role* in the information lifecycle with only some types of other stakeholders well served.

As an example, DDI [7] suggests some means to model information about funding but European funding bodies are likely to use their own information systems, many of them based on CERIF standard [6]. So the richness and expressivity of DDI, as well as the actual information systems based on it are in fact aimed at researchers in social science and data archivists, not at funders who are likely to have their own information systems based on other metadata standards, and not at other types of stakeholders in Business, Education, or researchers in other research domains.

We feel that it will be more productive to admit this natural attitude of the information management solutions and their owners to cater for only one or a few roles; it may be better to provide a reasonable means to model different roles and their activities on a granular level than try to capture an elusive information context in more and more complex versions of a comprehensive semantic model. If we take the existing records in a certain rich metadata format, this approach results in categorization and annotation of the entire metadata records with other metadata based on a smaller but semantically meaningful and universal information model – like our activity model.

Let us see how our core semantic model may serve DDI metadata categorization and annotation.[2] The analysis shows that one DDI record typically represents different types of research activity:

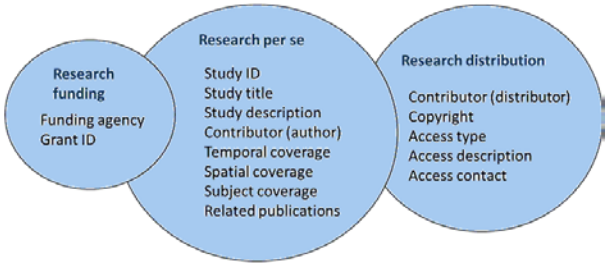

Figure 4. Research activities represented by a DDI record.

As we have identified different types of research activity, we can model them accordingly; we can also

[2] This approach was applied to DDI records harvested from the UK Data Archive and GESIS archive ([18], [13]) in the interests of the ENGAGE project [11] and was communicated in [3] as a prolegomenon to the generic model that we are presenting now.

identify specific Actors (Funding Agency, Author, Distributor), activity Outputs (Publication, Dataset), Scopes (Spatial Coverage, Subject Coverage) and Conditions (Copyright, Access Terms). Different granular activities will be modeled then with different amount of detail but we can enrich them with data from other information systems: for research funding – through funding agency portals, for research – through the project and the individual investigators' Web pages. This information enrichment should ideally be done by the Actors of the respective Activities (Funding, Research per se, Distribution) as they best understand the information context and the semantics of their business.

Our activity model then should allow curating the data and data context (metadata) in a distributed manner, and the combination of granular activities in sensible information context chains. This should eventually give us a more dispersed but a more complete description of the research discourse for a particular Study – more complete if compared to what the Data Archivist deemed valuable to capture and describe in a DDI record for the same. Our core model then serves as a "glue" to support the common information context and facilitate the interoperability of different digital curation frameworks that are operated by different Actors in support of their own Activities.

The existing well curated archives of DDI records can be considered then a valuable "fuel" to support the launch of the research discourse "Web" or "grid". The role-centric nodes of it will be performing their part of digital curation, with sharing its results via simple and commonly understandable semantic model that can be interpreted not only by data archivists or researchers in social science but by various stakeholders from other research domains, or business, or education, or policy making.

2.5 Conclusion

We outlined the motivation for why a simple model would be valuable for the semantic representation of a generic research lifecycle. We introduced the major aspects of the model, suggested an RDF manifestation for them and showed how the domain-agnostic requests might work for information discovery. We then considered a particular business case of applying the model to the existing rich metadata records in social science but there are more promising cases to consider.

One of the immediate candidates is facilities science with its CSMD metadata [5] that we already mentioned. The diverse business practices for using the existing mature data management solutions based on CSMD model [14] may become a barrier to the meaningful sharing of facilities science data as Linked Data. Our model then may be of help for the re-engineering of the existing data archives in spirit of Linked Data and Semantic Web principles, through semantic annotation of the CSMD metadata records (which may involve some decomposition, too, similarly to what we demonstrated for DDI metadata).

Another prospective area where we think our model may prove to be valuable is long-term digital preservation with its two well-known problems of the accountable data provenance and of the meaningful data representation for the future (and changing) community of data consumers. The ability of our model to combine individual data curation activities into the traceable chains of them, as well as its very focus on the Activity (with data being an artefact or footprint of it) may contribute to the satisfactory resolution of the data provenance problem. The model's data discovery capabilities based on standard information requests and profiles of them when it is enough for the User to be familiar with our basic semantic model in order to start the incremental knowledge discovery – may contribute to the meaningful data representation.

Also we find the multi-disciplinary and *distributed* curation, discovery and re-use of the research information to be in high demand; it is already in the agenda of a few actual European projects (see under [11], [12], [16]) and it is reasonable to expect more of them to come. The domain-agnostic nature of our model, as well as its very manageable core size and expandability where required let us hope for its application in some of the existing and future e-infrastructure initiatives.

3 Appendix: RDFS Plus manifestation of the activity model

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix am: <http://example.org/stuff/ActivityModel#> .

############### Core entities of Activity model ###############

# Comments are based on the Oxford dictionary, with some generalization or amendment where appropriate

am:Activity rdf:type rdfs:Class ;
        rdfs:label "Activity" ;
        rdfs:comment "Something that Actor does, or has done,
            or is going to do, or can do" .
am:Input rdf:type rdfs:Class ;
        rdfs:label "Activity Input" ;
        rdfs:comment "Something that is taken in or operated on
            by Activity" .
am:Output rdf:type rdfs:Class ;
        rdfs:label "Activity Output" ;
        rdfs:comment "Something that is intentially produced
            by Activity" .
am:Actor rdf:type rdfs:Class ;
        rdfs:label "Activity Actor" ;
        rdfs:comment "Something or somebody who participates
            in Activity" .
am:Effect rdf:type rdfs:Class ;
        rdfs:label "Activity Effect" ;
        rdfs:comment "Something that is a consequence
            of Activity" .
am:Condition rdf:type rdfs:Class ;
        rdfs:label "Activity Condition" ;
        rdfs:comment "Something that affects or supports
            Activity, or gives it a specific context" .
am:Scope rdf:type rdfs:Class ;
        rdfs:label "Activity Scope" ;
        rdfs:comment "Something that Activity is aimed at
            or deals with" .
```

```
########## Core properties of Activity model ##########

# am:hasInput or am:inputFor
# links Activity to its Input
am:hasInput owl:inverseOf am:inputFor .

# am:hasOutput or am:outputOf
# links Acttivity to its Output
am:hasOutput owl:inverseOf am:outputOf .

# am:hasActor or am:actorFor
# links Activity to its Actor
am:hasActor owl:inverseOf am:actorFor .

# am:hasEffect or am:effectOf
# links Activity to its Effect
am:hasEffect owl:inverseOf am:effectOf .

# am:hasCondition or am:ConditionFor
# links activity to its Condition
am:hasCondition owl:inverseOf am:ConditionFor .

# am:hasScope or am:ScopeOf
# links Activity to its Scope
am:hasScope owl:inverseOf am:scopeOf .
```

References

[1] Tim Berners-Lee. Open, Linked Data for a Global Community. A talk given on Gov 2.0 Expo, Washington, DC, 26 May 2010. http://www.gov2expo.com/gov2expo2010/public/schedule/detail/14247

[2] Juan Bicarregui, Vasily Bunakov, and Michael Wilson. PANdata international information infrastructure for synchrotrons: opportunity for collaboration. Presentation on the 19th Russian Synchrotron Radiation Conference (SR-2012), Novosibirsk, Russia, 25-28 June 2012. http://epubs.stfc.ac.uk/work-details?w=63074

[3] Vasily Bunakov. Semantic categorization of DDI metadata. Presentation on the 4th Annual European DDI User Conference (EDDI12), Bergen, Norway, 03-04 Dec 2012. http://epubs.stfc.ac.uk/work-details?w=64315

[4] Vasily Bunakov and Brian Matthews. Data curation framework for facilities science. In Proceedings of DATA 2013: the 2nd International Conference on Data Management Technologies and Applications, p.211-216, Reykjavík, Iceland, 29-31 July 2013.

[5] Brian Matthews et al., 2012. Model of the data continuum in Photon and Neutron Facilities. PaNdata ODI, Deliverable D6.1. http://pan-data.eu/sites/pan-data.eu/files/PaNdataODI-D6.1.pdf

[6] Common European Research Information Format. See under www.eurocris.org

[7] Data Documentation Initiative – Lifecycle Specification. http://www.ddialliance.org/Specification/DDI-Lifecycle/

[8] Semantic Statistics for Social, Behavioural, and Economic Sciences: Leveraging the DDI Model for the Web. Schloss Dagstuhl, September 11 – 16, 2011. http://www.dagstuhl.de/en/program/calendar/evhp/?semnr=11372

[9] DDI Lifecycle: Moving Forward. Schloss Dagstuhl, October 21 – 26, 2012. http://www.dagstuhl.de/en/program/calendar/evhp/?semnr=12432

[10] DARIAH-EU: Digital Research Infrastructure for the Arts and Humanities. http://www.dariah.eu/

[11] ENGAGE: An Infrastructure for Open, Linked Governmental Data Provision towards Research Communities and Citizens. http://www.engage-project.eu/

[12] EUDAT: European Data Infrastructure. http://www.eudat.eu/

[13] GESIS - Leibniz-Institut für Sozialwissenschaften. http://www.gesis.org/

[14] ICAT project. http://www.icatproject.org/

[15] Infrastructure for Integration in Structural Sciences (I2S2) Project. http://www.ukoln.ac.uk/projects/I2S2/

[16] PaNdata: Photon and Neutron Data Infrastructure. http://pan-data.eu/

[17] Structured Analysis and Design Technique. http://en.wikipedia.org/wiki/Structured_Analysis_and_Design_Technique

[18] UK Data Archive (for social sciences and humanities). http://data-archive.ac.uk/