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.

Scientific Methods Metadata for Provision of the Methods Reuse and Result Reproducibility

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

Data-intensive sciences are characterized by the constantly growing needs for specific data analysis methods intended for producing new knowledge related to the investigated areas. Development of new data analysis methods becomes a significant, inseparable part of research. Digital preservation, reuse and reproducibility of computer experiment results become inherent attributes of scientific discovery. The paper investigates metadata structure to be attached to the processes specifying or implementing scientific data analysis methods for their reuse and result

reproducibility. Process components and data are referred to the domain concepts and need to be supplied with the information about data provenance and quality. Specific test collections are needed to describe kinds of cases in which methods must behave in an anticipated way. Using the open myExperiment environment organizing and providing access to the collection of scientific workflows as an illustration, we demonstrate how the extension of its metadata could have allowed to organize the semantic search for methods relevant to a problem, to verify interoperability, reusability and reproducibility of processes implementing the methods.

.