Концептуальное моделирование предметных областей с интенсивным использованием данных

© Н. А. Скворцов © Л. А. Калиниченко © Д. Ю. Ковалев

Институт проблем информатики ФИЦ ИУ РАН, Москва

nskv@mail.ru leonidandk@gmail.com dm.kovalev@gmail.com

Аннотация

Исследования в различных предметных областях, особенно в направлениях естественных наук, связаны сегодня с обработкой больших объёмов данных наблюдений, экспериментов и моделирования. При организации исследований с интенсивным использованием данных целесообразно определять спецификации предметных областей, включающие определения понятий предметных областей средствами онтологий и абстрактное представление данных об объектах предметных областей и их поведении средствами концептуальных схем, разделяемых и поддерживаемых работающими в этих предметных областях сообществами. Исследовательские инфраструктуры опираются на спецификации предметных областей и предоставляют реализации методов, применимых над такими спецификациями, накапливаемых и развиваемых сообществами исследователей. Средства проведения экспериментов в инфраструктурах исследований также поддерживаются концептуальными спецификациями, которые обеспечивают основу для проведения измерений, изучения свойств сущностей предметной области, применения методов данной предметной области, описания и проверки гипотез. На примере предметной области астрономии показаны принципы построения концептуальных спецификаций и их использования при анализе данных.

Работа выполнена при частичной поддержке РФФИ (гранты 16-07-01028, 16-07-01162, 15-29-06045, 14-07-00548).

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

1 Введение

Исследовательские задачи критически зависят от растущих и дополняющих одна другую массивных коллекций данных, собираемых в результате наблюдений, экспериментов и моделирования. Одновременно растёт качество данных и соответственно глубина требуемого анализа данных. Подходы к исследованиям, при которых для решения задач производился выбор источников данных и формулирование задач в их терминах, стали трудоёмкими при множестве неоднородных источников данных и большом количестве способов их анализа. Если программы, реализующие решение задач анализа данных, зависят от конкретных источников данных, это препятствует масштабированию для неоднородных и массивных источников данных, накоплению реализаций методов анализа данных, их интероперабельности и повторному использованию в различных исследованиях [1].

От поиска и связывания источников данных для решения поставленных задач акцент исследований смещается в направлении анализа доступных массивных коллекций данных для нахождения новых знаний в предметной области исследования [2]. Разрабатываются научные методы оценки характеристик объектов по наблюдаемым параметрам, методы обобщения, классификации, выявления и исследования интересующих сущностей и явлений, средства генерации и проверки научных гипотез, специализированные процедуры в определённых направлениях науки, обеспечивается их автоматизированное применение над данными массивных коллекций и доступность для сообществ, работающих в инфраструктурах исследований.

Для разностороннего изучения конкретных типов сущностей реального мира оказывается важным совместное использование средств исследований и концептуальных спецификаций, определяющих как семантику сущностей и явлений в предметной области, так и семантику применяемых в ней

методов. Поэтому одной из задач сообществ, занимающихся исследованиями в определённой предметной области, является концептуализация предметной области для построения таких спецификаций и связывания с ними данных, реализаций методов и процессов.

Для обеспечения науки методами и средствами, применимыми к объектам предметных областей, в работе предлагается подход к концептуализации предметных областей для их исследования. Он опирается на явное описание семантики сущностей и процессов при формулировании постановок и реализаций алгоритмов решения научных задач, обеспечивая их семантическое соответствие спецификациям предметной области. В свою очередь, различные источники данных, в том числе научных данных, семантически отображаются в концептуальные спецификации предметной области исследования. Аналитические задачи формулируются также в терминах концептуальных спецификаций предметной области и решаются с использованием отображённых в них данных и методов.

В настоящей работе показано, каким образом концептуальные спецификации предметных областей, поддерживаемые заинтересованными сообществами, могут быть использованы для организации исследований. В статье для этого используются концептуальные спецификации предметных областей из области астрономии. Следующий раздел посвящён принципам определения спецификаций предметных областей исследований. В разделе 3 описаны подходы к накоплению научных методов и перспективы построения инфраструктур исследований на основе коллекций методов. Раздел 4 посвящён использованию спецификаций предметных областей для проведения экспериментов, описания и проверки научных гипотез, организации потоков работ в инфраструктурах исследовательских сообществ для манипулирования данными и методами при проведении экспериментов.

2 Средства спецификации предметных областей

Процесс концептуализации предметных областей, в первую очередь, предполагает разработку онтологий в исследовательских сообществах для формализации и систематизации знаний о характерных для этих областей сущностях и явлениях. Члены сообществ действуют в рамках онтологического обязательства, определённого такими онтологиями, то есть используют понятия предметных областей непротиворечивым образом по отношению к теориям, специфицируемым онтологиями. Для обеспечения такого подхода важна автоматизация контроля непротиворечивости результатов действий при любых манипуляциях понятиями предметной области.

Концептуальные определения предметных областей для проведения исследований включают следующие описания:

- понятия сущностей, фигурирующих в предметной области в качестве исследуемых или связанных с ними;
- понятия, определяющие характеристики и поведение объектов предметной области;
- понятия, соответствующие научным методам, корреляциям, существующим в предметной области исследования;
- понятия, определяющие подходы к наблюдению объектов и моделированию сущностей предметной области, проведению научных экспериментов.

Языковые средства представления формальных онтологий включают понятия, отношения понятий и ограничения, связанные с понятиями, обычно выраженные в подмножестве логики предикатов, отображаемом в некоторую дескриптивную логику или другие формальные модели. Скалярные типы данных в онтологиях предпочтительно не использовать, так как они отражают некоторые отношения, которые на онтологическом уровне лучше описывать явно для однозначной интерпретации понятий. Также в онтологиях традиционно не используются средства спецификации методов. Однако ограничения, связанные с поведением объектов предметной области, необходимо специфицировать и в онтологии. Это делается посредством определения понятий, соответствующих разного рода корреляциям характеристик сущностей и процессам.

Процесс концептуализации предметной области помимо определения понятий включает разработку концептуальных схем предметных областей, отличающихся от онтологий, в первую очередь, своим назначением [24]. Они определяют не просто понятия предметной области, а структуры представления информации об объектах предметных областей и спецификации поведения для манипулирования объектами. Однако, если разработаны онтологии, то концептуальные схемы составляются согласно знаниям об сущностях, зафиксированным в этих онтологиях. Принципы составления концептуальных схем предметных областей на основе определений онтологий описаны в [3]. Языковые средства спецификации концептуальных схем включают определения абстрактных типов данных, представляющие информацию о состоянии объектов и характеризующихся наборами атрибутов, значения которых соответствуют определённым типам данных от простых скалярных до объектных типов и ассоциаций. С типами и атрибутами типов могут быть связаны метаданные, определяющие их собственные характеристики. Множества однотипных объектов могут составлять классы. Поведение объектов предметной области выражается методами типов.

Любые структуры или информационные объекты целесообразно сопровождать метаинформацией о том, с какими понятиями онтологии они связаны, чтобы фиксировать их семантику и систематизировать в соответствии с ней ресурсы, имеющиеся в арсенале исследователей определённой предметной области.

Формальность спецификации онтологий и концептуальных схем принципиально важна для обеспечения семантической интеграции информационных ресурсов и воспроизводимости программ над спецификациями предметной области. Без доказательного подхода использование спецификаций предметной области мало отличается от умозрительного связывания элементов схем при интеграции ресурсов. Формального обоснования в концептуальном подходе требуют такие задачи, как, например:

- проверка внутренней непротиворечивости спецификаций онтологий и концептуальных схем предметной области:
- контроль интероперабельности совмещаемых или замещаемых спецификаций;
- проверка соответствия разрабатываемых спецификаций концептуальных схем знаниям, отражённым в онтологии;
- обнаружение спецификаций информационных ресурсов, семантически соответствующих спецификациям предметной области;
- проверка соответствия используемых информационных ресурсов спецификациям предметной области.

От выбранного формализма языка спецификации зависит возможность применения средств автоматического вывода. В частности, дескриптивные логики используются в качестве основ диалектов языка онтологий OWL. Для спецификаций, приводимых к дескриптивным логикам, перечисленные выше задачи разрешимы.

Спецификации в моделях, основанных на логиках, целесообразно представлять в унифицированном виде, в частности, в диалектах языка RIF [17]. Помимо прочего, язык RIF может использоваться для выражения правил над спецификациями на языке OWL, что позволяет определять формальные спецификации поведения объектов предметной области и алгоритмы решения задач напрямую над онтологиями OWL. В мультидиалектной архитектуре в зависимости от используемых диалектов RIF для рассуждений над спецификациями используются соответствующие им системы вывода [23].

Для языков, выразимых в логике предикатов первого порядка, те же задачи могут быть решены в интерактивном режиме при помощи доказательства уточнения спецификаций программ [18]. Уточняющая спецификация может быть использована вместо уточняемой. В частности, при разработке спецификаций концептуальных схем предметных областей на основе онтологии необходимо, чтобы онтология уточнялась спецификациями схем. Для обоснования этого язык онтологий и язык концептуальных схем должны быть отображены в язык абстрактных машин системы вывода B, обеспечивающей доказательство уточнения [19-21]. При этом понятия, определяющие зависимости и процессы, отображаются в операции, которые должны уточняться операциями типов в концептуальной схеме. Данные о том, какие элементы схемы сформированы в соответствии с понятиями онтологии, сохраняются в схеме в качестве метаданных. Таким образом, явно специфицируется семантика элементов схем с точки зрения понятий предметной области.

Онтологии и концептуальные схемы предметных областей разрабатываются и поддерживаются сообществами, работающими в этих областях, таким образом, чтобы быть достаточными для нужд научных групп. Средства и состав концептуальных спецификаций предметных областей в сообществе направлены на семантическую интероперабельность взаимодействующих компонентов, повторное использование информационных ресурсов и воспроизводимость программ за счёт привязки к семантике предметной области. Поэтому как на уровне онтологий, так и на уровне концептуальных схем предъявляются высокие требования к полноте и формальности спецификаций.

Оценивая использование концептуальных спецификаций на примере области астрономии, необходимо отметить, что в рамках альянса Международной виртуальной обсерватории разрабатываются соответствующие стандарты. Известны онтологии [4, 5], однако они созданы на основе тезаурусов и не содержат многих существенных понятий и отношений, которые необходимы для работы исследователей, не отражают ограничений состояния и поведения объектов, явлений и научных экспериментов в предметной области. Нет хорошо формализованных онтологий, направленных на логический вывод.

К концептуальным спецификациям в астрономии относятся также разрабатываемые стандарты концептуальных схем наиболее общих областей, которые затрагиваются практически в каждой астрономической задаче, в частности:

- Space-Time Coordinate Metadata [6] – схема свойств различных систем координат;
- Photometry Data Model [7] – схема и формат сериализации фотометрической информации, определяющий также функции калибровки и преобразования между разными фотометрическими системами;
- VOEvent [8] – схема описания наблюдаемых объектов и астрофизических явлений, включающая идентификацию объектов, наблюдателей, место, время и средства наблюдения.

Схемы, как и онтологии Международной виртуальной обсерватории, не описывают объекты и научные методы специфических областей и, в основном, не включают ограничений целостности и спецификаций поведения объектов.

Концептуализация необходима не только в наиболее общих областях, затрагиваемых астрономией, таких как фотометрия и спектроскопия, но и в областях, представляющих интерес для более узких кругов исследователей, а также на границе между областями, где чаще всего возникает сотрудничество научных коллективов и повторное использование результатов исследований. При этом важно описание как объектов исследования, так и методов исследования и проведения экспериментов в таких областях. Прототип[1] разработанной авторами статьи онтологии в области астрономии, определяющей некоторые специфические области наряду с общеупотребимыми понятиями, представлен на языке OWL [9].

Модульная структура онтологии, по сути, описывает взаимодействующие предметные области в рамках астрономии. Она включает области, определяемые разными объектами исследования, методами наблюдения и моделирования, подходами к исследовательскому процессу в целом. По мере расширения затрагиваемых областей и круга задач, решаемых взаимодействующими группами исследователей, в онтологии развиваются различные модули.

Разработана онтология для спецификации научных экспериментов, формирующая междисциплинарные базисные понятия. Она включает: 1) онтологию характеристик измерений объектов исследования, включающую такие понятия как единицы измерений, погрешности, законы распределения значений и другие; 2) онтологию взаимозависимостей измерений объектов, необходимую для введения понятийного аппарата для спецификации поведения объектов и включающую понятие корреляции измерений и его подпонятия, определяющие понятия функции, метода, закона, гипотезы и другие.

Рассмотрим некоторые спецификации на языке OWL. В части онтологии, определяющей понятия, используемые для проведения научных экспериментов, определены несколько модулей. Среди них модуль, содержащий базовые понятия, относящиеся к измерениям параметров исследуемых объектов, включает понятие измерения (Measurement), связываемое с объектом исследования (AstrObject) отношением isMeasurementOf, понятия значений параметров, единиц измерений, точности измерений, характеризуемой статистической и систематической ошибками.

```
Class(Measurement
  restriction(hasValue
    allValuesFrom(Value))
  restriction(hasUnit
    allValuesFrom(MeasurementUnit))
  restriction(hasError
    allValuesFrom(MeasurementError)
    maxCardinality(1))
  restriction(isMeasurementOf
    allValuesFrom(AstrObject)
    maxCardinality(1)))
Class(MeasurementUnit
  restriction(hasScaleFactor
    allValuesFrom(ScaleFactor)
    maxCardinality(1))
  restriction(hasProjection
    allValuesFrom(ScaleProjection)
    maxCardinality(1)))
Class(MeasurementError
  restriction(isErrorOf
    allValuesFrom(Measurement)
    maxCardinality(1)))
Class(StatisticalError
  partial MeasurementError)
Class(SystematicError
  partial MeasurementError)
Class(Value
  restriction(isValueOf
    allValuesFrom(Measurement)
    maxCardinality(1)))
ObjectProperty(isValueOf
  domain(Value)
  range(Measurement)
  inverseOf(hasValue))
ObjectProperty(isMeasurementOf
  domain(Measurement)
  range(AstrObject)
  inverseOf(hasMeasurement))
ObjectProperty(isErrorOf
  domain(MeasurementError)
  range(Measurement)
  inverseOf(hasError))
```

Модуль астрономических объектов определяет понятия, связанные с характеристиками, общими для произвольных астрономических объектов. Спецификация понятия астрономического объекта (AstrObject) включает связи с другими понятиями онтологии: его координатами, измерениями разного рода физических параметров, связью с составными объектами, к которым данный объект принадлежит в качестве компонента, и другими:

```
Class(AstrObject
  restriction(hasIdentifier
    allValuesFrom(Identifier))
  restriction(hasCoordinate
    allValuesFrom(Coordinate))
  restriction(inEpoch
    allValuesFrom(Epoch)
    maxCardinality(1))
  restriction(hasMeasurement
    allValuesFrom(Measurement))
  restriction(hasMorphology
    allValuesFrom(Morphology)
    maxCardinality(1))
  restriction(hasProcess
    allValuesFrom(Process))
  restriction(isComponentOf
    allValuesFrom(CompoundObject)))
```

Онтологический модуль, описывающий предметную область звёзд, включает понятие звёздного объекта (StellarObject) как точечной сущности в Галактике, самостоятельного или являющегося компонентом составного объекта,

[1] http://ontology.ipi.ac.ru/ontologies/astront/

понятие звезды (Star) как одиночного звёздного объекта, понятие кратной звезды как звёздного объекта, состоящего из компонентов, а также ряд специфических понятий характеристик звёзд:

```
Class(StellarObject
  partial AstrObject
  restriction(hasMorphology
    hasValue(PointObject)))
Class(Star
  partial StellarObject)
```

Конкретные виды измерений определяются как подпонятия понятия Measurement в специализированных модулях онтологии, в которых они используются. Модуль астрофизических параметров астрономических объектов содержит общие физические характеристики объектов, такие как температура, масса, размеры, светимость. В частности, масса является общей характеристикой астрономических объектов:

```
Class(Mass
  partial Measurement)
```

Представим себе понятие массы звезды (StarMass), являющееся подпонятием понятия масса (Mass). Оно использует понятия разных модулей, ограничивая тип описываемых сущностей как звёзды и определяя в качестве единицы измерения массу Солнца.

```
Class(StarMass
  partial Mass
  restriction(isParameterOf
    allValuesFrom(Star))
  restriction(hasUnit
    hasValue(SunMass)))
```

При переходе от онтологии к концептуальной схеме необходимо сформировать структуры для представления информации об объектах предметной области. В мультидиалектной архитектуре помимо спецификаций OWL используется язык СИНТЕЗ [10] для реализации на основе применения предметных посредников. В то время как OWL является разрешимым языком для задачи включения, для языка СИНТЕЗ отработано доказательство уточнения спецификаций. В представление на языке СИНТЕЗ также могут быть отображены спецификации в диалекте RIF BLD. Также разработано отображение языка OWL в язык СИНТЕЗ [11].

Пример спецификации концептуальной схемы на языке СИНТЕЗ, построенной в соответствии с онтологической спецификацией, определяет структуру представления информации о звёздах с атрибутом, хранящим массу звезды и метаданными, определяющими единицу её измерения в массах Солнца:

```
{ Star;
  in: type;
  crd: Coordinate;
  mass: Float;
    metaslot
      in: measurement;
      hasUnit: SunMass;
    end
}
```

Спецификации концептуальных схем требуют также определения методов и функций. Такие спецификации формируются в схеме на основе понятий онтологии, описывающих зависимости характеристик объектов, а также понятий процессов. Они рассматриваются в следующем разделе.

3 Организация коллекций научных данных и методов

Сегодня активно развиваются библиотеки методов в специализированных областях исследований в астрономии, в инфраструктурах совместных исследований, где помимо данных в доступ исследовательскому сообществу предоставляются всевозможные сервисы, а также средства их поиска и описания для правильного использования.

Одной из первых систем, предоставляющих технологии для работы сообществ в области астрономии была сеть AstroGrid [12]. Она представляла собой инфраструктуру для решения задач виртуальной обсерватории и состояла из множества узлов, содержащих всевозможные сервисы и ресурсы. Архитектура AstroGrid включала реестр, представляющий собой коллекцию метаданных, описывающих ресурсы, которые могут использоваться при решении задач. Это позволяло организовать поиск доступных коллекций данных и методов по метаданным. Проект был закрыт, в первую очередь, по причине медленного развития сети. Организация узлов сети оказалась сложной для широкого распространения в астрономическом сообществе.

Проект WF4Ever [13] направлен на сохранение результатов научных исследований над данными, с этой целью разработаны средства курирования объектов исследования как комплексных объектов, включающих документы, данные, сервисы, потоки работ. В рамках этого же проекта развивается библиотека сервисов [14], которые обеспечивают доступ к существующим астрономическим веб-сервисам и к данным каталогов, преобразование между разными стандартными представлениями и манипулирование таблицами при соединении разных источников данных, и используются в качестве элементов потоков работ. Библиотека получила признание научного сообщества, в первую очередь, за счёт простоты построения процессов обработки данных без программирования.

В проекте EUDAT [22] ставится задача построения инфраструктуры доступа к научным данным. Семантические подходы к её организации включают ведение репозиториев словарей, включающих термины широкого круга научных областей. Помимо словарей общего назначения, определяющих такие атрибуты как название, авторы, научная дисциплина, определяются иерархии терминов, именующих научные дисциплины, научные методы и объекты. Специфические для предметных областей словари разрабатываются научными сообществами. Эти описания используются для организации информационно-поисковой системы, обеспечивающей поиск

релевантных задаче сервисов. Поиск может производиться одновременно по терминам разных словарей на пересечении исследований разных научных сообществ. Европейская виртуальная обсерватория представлена в проекте как одна из областей исследования, и её решения сравниваются с подходами EUDAT.

В ближайшей перспективе в области астрономии появятся источники данных, объёмы которых намного превышают сегодняшние. Такие проекты как широкоугольный телескоп LSST и космическая обсерватория Gaia будут генерировать потоковые данные наблюдений. Для их обработки заранее готовятся каналы передачи данных для их локализации в местах исследований, решаются вопросы доступа к данным различных научных учреждений, а также разрабатываются общедоступные средства обработки данных и средства их эффективного поиска [15].

Для эффективного взаимодействия внутри сообщества, имеющего доступ к данным, и во избежание появления множество разрозненных работ кооперация исследователей в подобных проектах должна основываться на обеспечении доступа к разработке планов исследований, специализированным методам и результатам анализа данных. Таким образом, помимо накопления данных, необходимо накопление доступных методов, алгоритмов и инструментов обработки, готовых к применению над большими массивами данных. Концептуализация предметной области в рамках сообщества и семантические подходы позволяют систематизировать методы предметной области в исследовательских инфраструктурах. И научные данные, и научные методы связываются со спецификациями предметных областей, к которым они относятся.

Во всякой предметной области накапливаются знания и законы предметной области, специфические методы, направленные на определённые виды анализа сущностей, фигурирующих в предметной области. Помимо этого, должен быть доступен широкий круг аналитических методов и инструментов общего назначения, применяемых вне рамок специфической предметной области. К таким методам относятся, например, численные, статистические методы, методы машинного обучения и другие.

В рамках онтологических моделей, традиционно не имеющих средств спецификации методов, концептуализация поведения объектов может определяться понятиями, отражающими зависимые характеристики объектов и процессы. Одним из модулей разрабатываемой онтологии является модуль, определяющий взаимозависимости измерений. Под понятием корреляции (Correlation) подразумевается корреляция определённых параметров измерения у объектов предметной области:

```
Class(Correlation
  restriction(isCorrelationOf
    allValuesFrom(Measurement))
  restriction(hasRegression
    allValuesFrom(RegressionFunction))
  restriction(hasRMSDeviation
    allValuesFrom(RMSDeviation))
  restriction(isCausal
    allValuesFrom(TruthValue)))
Class(Hypothesis
  partial Correlation
  restriction(explains
    allValuesFrom(Phenomenon))
  restriction(derivedFrom
    allValuesFrom(Hypothesis))
  restriction(competesWith
    allValuesFrom(Hypothesis))
  restriction(hasProbability
    allValuesFrom(Probability))
  restriction(hasPValue
    allValuesFrom(Probability))
  restriction(hasQuality
    allValuesFrom(TruthValue)))
Class(Law
  partial Hypothesis
  Restriction(hasQuality
    hasValue(True)))
```

Посредством понятий на уровне онтологии декларативно описываются научные методы, гипотезы, законы, модели, процессы, эксперименты, связанные с характеристиками объектов предметной области. Понятие гипотезы определяется как разновидность статистически подтверждаемых корреляций. Для статистического подтверждения гипотез, с одной стороны, используется моделирование, обеспечивающее их математическое описание, а с другой стороны, – эксперимент для сравнения результатов моделирования с данными наблюдения объектов исследуемых объектов.

Понятия научных методов, законов и гипотез определяются как подпонятия корреляции измерений с указанием ограничений конкретных зависимых величин. Их понятийное описание не зависит от конкретных реализаций и представлений, будь то таблиц значений или коэффициентов, точных математических формул, функций распределения, программ или других возможных способов описания.

Рассмотрим спецификации концептуализации гипотезы начальной функции масс в составе Безансонской модели Галактики [25] на основе онтологических спецификаций предметной области, приведённых выше. Эта гипотеза связана с предположением о достаточно постоянном распределении звёзд разной массы в некотором ограниченном объёме пространства Галактики. Другими словами, гипотеза предполагает зависимость количества звёзд от их массы в фиксированном объёме пространства, которому принадлежат эти звёзды:

```
Class(InitialMassFunction
  partial Hypothesis
  restriction(isCorrelationOf
    ObjectSomeValuesFrom(StarMass))
  restriction(isCorrelationOf
    ObjectSomeValuesFrom(
      intersectionOf(
        Quantity
        restriction(hasElement
          allValuesFrom(Star))))))
```

На основе онтологии коллекции различных реализаций методов могут быть систематизированы по различным признакам, соответствующим понятиям онтологий: исследуемым объектам, характеристикам объектов, свойствам, зависимым от данной характеристики, известным методам и гипотезам и другим. Соответственно, по любому из таких признаков исследователями может производиться поиск существующих реализаций научных методов для их повторного использования.

4 Средства проведения научных экспериментов

Применение методов, собранных в коллекции, при исследованиях в предметной области происходит в соответствии с определёнными сценариями. Так, автоматизированный запуск анализа данных может происходить при появлении данных с определёнными свойствами или определённого типа для обогащения данных об объектах определёнными характеристиками, которые в свою очередь могут использоваться для дальнейших исследований.

Методика исследования обычно состоит из определённых шагов, включающих очистку и анализ данных, построение научных гипотез, моделирование в соответствии с гипотезами и проверку моделей на данных наблюдений. Эксперименты над данными формулируются на основе создания новых методов и повторного использования существующих реализаций методов в спецификациях потоков работ.

Инфраструктуры поддержки научных исследований, помимо возможности использования коллекций данных и реализаций методов в определённых предметных областях, должны содержать средства проведения научных экспериментов. В частности, это касается возможности формулирования и проверки научных гипотез.

Использование концептуальных спецификаций при формулировании и тестировании гипотез даёт те же преимущества, что и при управлении коллекциями методов и решении научных задач над ними.

На уровне концептуальных спецификаций понятиям методов и законов и гипотез приводятся в соответствие методы и правила. Спецификации ограничений понятий зависимых величин уточняются предусловиями и постусловиями методов.

Так на основе знаний, специфицированных в понятии, описывающем гипотезу начальной функции масс могут быть созданы абстрактные типы данных концептуальной схемы для моделирования и проверки гипотез. Тип будет включать определение интерфейса функции с параметрами, соответствующими отношениям в понятии зависимости.

```
{ IMF;
  in: type;
  supertype: Hypothesis;
  draw_mass: { in: function;
    params: { +mass/Real, –quantity/Real } };
}
```

Если в онтологической спецификации определялось направление зависимости (функция), в соответствии с ним определяются и входные и выходные параметры. Иначе выбор направления определяется нуждами решаемой задачи. В предусловии и постусловии функции определяются ограничения, соответствующие онтологическим определениям. Ограничения в онтологии, связанные с измерениями, единицами измерений, точностью измерений, помогают сформировать структуру метаданных для сопровождения измерений в концептуальной схеме. Таким образом, концептуальные спецификации предметной области используются для моделирования закономерностей Галактики и исследования моделей.

Представленный абстрактный тип данных может быть реализован разными способами для организации проверки гипотезы. Существующие реализации моделей, соответствующих гипотезе, могут быть найдены в коллекции методов по онтологическим описаниям. Для конкретной реализации определяется подтип данной спецификации.

С другой стороны, та же спецификация типа используется для проверки гипотезы на данных экспериментов. Для этого используются данные из множества источников, которые интегрированы в концептуальную схему предметной области. Подтип вышеприведённой спецификации строится таким образом, чтобы по входным данным получать данные наблюдения, соответствующие выходным параметрам. Проверка гипотезы производится с помощью сравнения результатов моделирования с результатами, полученными на данных реальных источников.

Реализация моделей и экспериментов может использовать доступные методы общего назначения, такие как методы машинного обучения или численные методы, однако спецификации онтологий и схем, принятых в сообществе, от них не зависят.

Эффективность исследований, проводимых сообществом предметной области, зависит не только от доступности данных наблюдения, реализаций методов и моделей, но также от планирования экспериментов, в котором учитываются онтологические знания об изучаемых объектах и взаимозависимостях их характеристик (гипотезах и законах).

Концептуальные спецификации могут быть использоваться при генерации гипотез. Генерация на основе поиска корреляций в данных требует проверки семантического соответствия коррелирующих параметров объектов. Не связанные друг с другом хотя бы опосредованно параметры в спецификациях понятий с меньшей долей вероятности рассматриваются как коррелирующие.

При взаимодействии разных гипотез в одной модели взаимное влияние их параметров, участвующих в гипотезах, может быть учтено на основе знаний онтологии о зависимости друг от друга разных измерений. Эти зависимости могут быть исследованы и их реализации найдены посредством семантического поиска с использованием онтологии. Ограничения концептуальных схем при этом могут гарантировать согласованность модели.

Для моделирования и проверки гипотез над концептуальными схемами разрабатываются потоки работ, которые реализуют процесс моделирования, проверки гипотез, их корректировки для подбора параметров моделей, наилучшим образом повторяющих результаты, полученные на реальных данных.

Заключение

В статье рассмотрены вопросы концептуализации предметных областей для организации научных исследований над данными. Развитие инфраструктур поддержки научных исследований, в основе которых лежат концептуальные спецификации предметных областей, развиваемые и поддерживаемые сообществами, работающими в этих областях, позволяет избежать зависимости программ от структуры источников данных, обеспечить интероперабельность различных методов при совместной работе, повысить надёжность результатов за счёт использования формальных непротиворечивых спецификаций. Рассмотрены возможности концептуального анализа предметной области для формализации научных гипотез и их тестирования на основе данных наблюдений.

Литература

[1] Д. О. Брюхов, А. Е. Вовченко, Л. А. Калиниченко. Поддержка повторного использования спецификаций потоков работ за счет обеспечения их независимости от конкретных коллекций данных и сервисов // Всероссийская конференция «Электронные библиотеки» RCDL 2013. – CEUR Workshop Proceedings, 2013. – Т. 1108. – С. 61-69.

[2] The Fourth Paradigm: Data-Intensive Scientific Discovery. T. Hey, et al (Eds). – Microsoft Research. – Redmond, 2009.

[3] А. Е. Вовченко, В. Н. Захаров, Л. А. Калиниченко и др. От спецификаций требований к концептуальной схеме // Труды 12-й Всероссийской научной конференции Электронные библиотеки: перспективные методы и технологии, электронные коллекции RCDL 2010. – Казань: КФУ, 2010. – С. 375-381.

[4] Ontology of Astronomical Object Types. Version 1.20. – IVOA, 2009. – URL: http://www.ivoa.net/documents/latest /AstrObjectOntology.html

[5] IVOAO Ontology. – University of Maryland, 2010. – URL: http://www.astro.umd.edu/~eshaya/astro-onto/

[6] Space-Time Coordinate Metadata for the Virtual Observatory
Version 1.33. – IVOA, 2011. – URL: http://www.ivoa.net/documents/ latest/STC.html

[7] IVOA Photometry Data Model. Version 1.0. – IVOA, 2013. – URL: http://www.ivoa.net/documents/PHOTDM/

[8] Sky Event Reporting Metadata (VOEvent). Version 2.0. – IVOA, 2011. – URL: http://www.ivoa.net/Documents/VOEvent/

[9] OWL 2 Web Ontology Language. Document Overview (Second Edition).} -- W3C, 2012. -- URL: http://www.w3.org/TR/owl-overview/

[10] L. A. Kalinichenko, S. A. Stupnikov, D. O. Martynov. SYNTHESIS: a Language for Canonical Information Modeling and Mediator Definition for Problem Solving in Heterogeneous Information Resource Environments. Moscow: IPI RAN, 2007. – 171 p.

[11] L. A. Kalinichenko, S. A. Stupnikov. OWL as Yet Another Data Model to be Integrated. Advances in Databases and Information Systems: Proc. II of the 15th East-European Conference. – Vienna: Austrian Computer Society, 2011. – P. 178-189.

[12] AstroGrid. – URL: http://www.astrogrid.org/

[13] K. Belhajjame, et al. Workflow-Centric Research Objects: A First Class Citizen in the Scholarly Discourse // ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web (SePublica2012). – Heraklion, 2012.

[14] N. A. Walton, et al. Taverna and workflows in the virtual observatory // Astronomical Data Analysis Software and Systems ASP Conference Series. – Vol. 394. – 2007. – P. 309.

[15] M. Luric, T. Tysoc. LSST Data Management: Entering the Era of Petascale Optical // Astronomy. Highlights of Astronomy. – Vol. 16. – 2015. – P. 675.

[16] N. A. Skvortsov, et al. Conceptual approach to astronomical problems // Astrophysical Bulletin. – Vol. 71, No. 1. – Springer, 2016.

[17] Rule interchange format: The framework // Web Reasoning and Rule Systems: 2nd Conference (International) Proceedings, LNCS 5348. – Berlin–Heidelberg: Springer Verlag, 2008. – P. 1-11.

[18] Abrial J.R. The B Book - Assigning Programs to Meanings. - Cambridge: Cambridge University Press, 1996.

[19] Н. А. Скворцов. Применение уточнения понятий в решении задач манипулирования онтологиями // RCDL'2007. – Переславль-Залесский: УГП, 2007. – С.225-229.

[20] Н. А. Скворцов. Использование системы интерактивного доказательства для

отображения онтологий // RCDL'2006. – Ярославль: ЯрГУ, 2006. – С. 65-69.

[21] С. А. Ступников. Отображение спецификаций, выраженных средствами ядра канонической модели, в язык AMN // Системы и средства информатики: Спец. вып. Формальные методы и модели в композиционных инфраструктурах распределенных информационных систем. Под ред. И. А. Соколова. М.: ИПИ РАН, 2005. С. 69-95.

[22] H. Schentz, Y. le Franc. Building a semantic repository using B2SHARE // EUDAT 3rd Conference. – 2014.

[23] L. A. Kalinichenko, S. A. Stupnikov, E. A. Vovchenko, D. Y. Kovalev. Rule-based Multi-dialect Infrastructure for Conceptual Problem Solving over Heterogeneous Distributed Information Resources // Advances in Intelligent Systems and Computing. – Springer, 2013. – V. 241. – P. 61-68.

[24] М. Р. Когаловский, Л. А. Калиниченко. Концептуальное моделирование в технологиях баз данных и онтологические модели. // Онтологическое моделирование: состояние и направления исследований и применения. - М. ИПИ РАН, 2008.

[25] A. C. Robin, C. Reylé, S. Derrière and S. Picaud. A synthetic view on structure and evolution of the Milky Way, 2003, Astron. Astrophys., 409:523 ADS

Conceptual modeling of subject domains in data intensive research

Nikolay A. Skvortsov, Leonid A. Kalinichenko, Dmitry Yu. Kovalev

Nowadays research of various scopes especially in natural sciences requires manipulation of big volumes of data generated by observation, experiments and modeling. Organization of data-intensive research assumes definition of domain specifications including concepts (specified by ontologies) and formal representation of data describing domain objects and their behavior (using conceptual schemes), shared and maintained by communities working in the respective domains. Research infrastructures are based on domain specifications and provide methods applied to such specifications, collected and developed by research communities. Tools for organizing experiments in research infrastructures are also supported by conceptual specifications of measuring and investigating object properties, applying of research methods, describing and testing of hypotheses. Astronomy as a sample data intensive domain (DID) is chosen to demonstrate building of conceptual specifications and usage of them for data analysis.