Подход к поиску потоков работ по метаданным

© Н.А. Скворцов
Институт проблем информатики РАН
nskv@ipi.ac.ru

Аннотация

Работа посвящена методам поиска реализаций потоков работ и их компонентов с целью повторного использования по спецификациям метаданных. Для спецификации потоков работ используются диалекты языка правил RIF, метаданные формулируются как аннотации RIF. Метаданные, необходимые для обеспечения повторного использования потоков работ, применяются в различных задачах, возникающих во время разработки потоков работ. В статье демонстрируются методы спецификации метаданных и семантического поиска потоков работ по ним.

1 Введение

Необходимость обработки больших объёмов данных и расширение направлений их обработки при исследованиях в науках с интенсивным использованием данных заставляет подходить к множеству средств обработки данных как к коллекциям научных методов, которые могут быть повторно используемы в различных задачах. Для организации обработки данных становится целесообразно разрабатывать потоки работ, которые представляют собой спецификации порядка обработки данных, обеспечивающего решение научных задач, и использовать существующие деятельности, сервисы, потоки работ.

Спецификации потоков работ в данном исследовании используют языки и технологии, применяемые в рамках Семантического веба. Основными средством спецификации потоков работ являются диалекты языка RIF [2] (Rule Interchange Format). Потоки работ специфицируются в мультидиалектной среде [6]. Деятельности потоков работ могут формулироваться в разных диалектах правил. Концептуальные схемы предметных областей, над которыми разрабатываются спецификации деятельностей, описываются средствами языка онтологий OWL 2.0 [1] и импортируются в спецификации потоков работ. Определенные в концептуальных схемах сущности могут использоваться в качестве предикатов в правилах. Оркестровка потока работ выражается посредством продукционных правил (в диалекте RIF PRD). В продукционных правилах могут использоваться предикаты, определённые в других диалектах при спецификации деятельностей. Для спецификации управляющих конструкций потоков работ определяется пространство имён со специальными предикатами:

- variable-definition и variable-value для организации потоков данных на основе переменных и их значений;
- parameter-definition и parameter-value для организации входных и выходных параметров потоков работ и значений параметров;
- end-of-task – индикатор завершения работы деятельности для организации последовательности, условий, разбиения, соединения и других шаблонов [9] потоков работ с помощью правил.

Например, следующая спецификация определяет шаблон разбиения по конъюнкции, в котором деятельности B и C выполняются одновременно после выполнения деятельности A:

```
If Not(External(wkfl:end-of-task(A)))
Then Do (Act(A)
  Assert(External(wkfl:end-of-task(A))))
If And(Not(External(wkfl:end-of-task(B)))
  External(wkfl:end-of-task(A)))
Then Do (Act(B)
  Assert(External(wkfl:end-of-task(B))))
If And(Not(External(wkfl:end-of-task(C)))
  External(wkfl:end-of-task(A)))
Then Do (Act(C)
  Assert(External(wkfl:end-of-task(C))))
```

Реализации потоков работ могут либо разрабатываться на основе спецификаций RIF при помощи трансляции правил в языки конкретных систем, работающих с определёнными диалектами правил, либо выбираться из существующих релевантных потоков работ, их фрагментов, отдельных деятельностей и сервисов.

Поиск релевантных потоков работ и их фрагментов производится в доступных коллекциях научных методов. Для возможности семантического поиска в таких коллекциях реализации потоков работ, помимо спецификации их структуры, сопровождаются определённым набором метаданных, несущих информацию о связи потоков

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

работ с понятиями предметной области, о качестве и происхождении используемых данных и методов. Состав необходимых метаданных был разработан ранее [13]. Эта информация обеспечивает не только возможность оценки потоков работ и их фрагментов с точки зрения структуры, но и учёт семантики предметной области и требований к качеству и надёжности работы научных методов.

Принципы семантического поиска подходящих реализаций потоков работ и их фрагментов на основе метаданных являются предметом исследования данной статьи. В следующем разделе описаны принципы связывания метаданных со спецификациями потоков работ на правилах. Затем приведён обзор методов поиска потоков работ. Последующие разделы рассматривают сценарии и методы семантического поиска релевантных потоков работ.

2 Связывание метаданных с потоками работ

Спецификации RIF несут формальную семантику правил, не позволяющую определять что-либо помимо правил в заданном диалекте. Для связывания со спецификациями дополнительной информации в языке предусмотрен механизм аннотирования. Аннотации могут сопровождать любой класс конструкций RIF в спецификациях правил. Они определяются как фреймы с наборами свойств этих конструкций, которые должны быть сохранены при любых манипуляциях спецификациями, но не добавляют семантики с точки зрения правил. Поэтому при реализации потоков работ спецификации метаданных игнорируются. Тем не менее, они могут обладать семантикой, не зависимой от правил.

Обычно аннотации в RIF определяются в терминах специализированного словаря, специфицирующего набор предопределённых свойств. Состав метаданных в настоящем исследовании не ограничивается набором свойств, а включает в себя более развитые описания. В качестве словарей метаданных используются онтологии предметных областей, а также онтологии, определяющие свойства элементов потоков работ в различных ракурсах рассмотрения, таких как качество и происхождение данных и методов.

Аннотации, которые определяют метаданные, целесообразно связывать со следующими элементами потоков работ, выраженных правилами:

- потоки работ в целом;
- входные и выходные параметры потоков работ;
- деятельности внутри потоков работ;
- входные и выходные параметры деятельностей;
- переменные, определяющие потоки данных;
- отдельные правила и группы правил, определяющие фрагменты потока работ;
- группы правил, определяющие шаблоны потоков работ [9].

Связывание метаданных с потоками работ и поиск релевантных элементов потоков работ далее рассмотрим на примере. В [6] описывается задача составления портфелей ценных бумаг, котировки которых не коррелируют друг с другом, и выбора лучшего из них по определённым критериям.

Для решения данной задачи разрабатываются спецификации потока работ, включающего:

- задачу поиска максимальных портфелей-кандидатов с независимыми друг от друга котировками бумаг;
- оценку бумаг, входящих в портфели, с точки зрения разных критериев, в частности, финансово-экономического и социального;
- оценка портфелей по соответствующим критериям как обобщение оценок бумаг, входящих в них;
- обобщение нескольких критериев оценки портфелей в общую оценку и выбор лучшего портфеля.

Для реализации потока работ используются данные об истории цен на бумаги, принадлежность компаний индексу S&P 500 (индекс оценивается на основе данных о капитализации пятисот крупных американских компаний), оценка соотношения доходности и риска, мониторинг тональности высказываний инвесторов об определённых бумагах. Оценка по разным критериям выполняется в потоке работ параллельными ветвями.

Для описания метаданных в терминах предметной области определяется онтология.

```
Class(Portfolio)
ObjectProperty(includesSecurity)
  ObjectPropertyDomain(includesSecurity Portfolio)
  ObjectPropertyRange(includesSecurity Security)

Class(Security)
ObjectProperty(hasIdentifier)
  FunctionalObjectProperty(hasIdentifier)
  ObjectPropertyDomain(hasIdentifier Security)
  ObjectPropertyRange(hasIdentifier Ticker)
ObjectProperty(listedIn)
  ObjectPropertyDomain(listedIn Security)
  ObjectPropertyRange(listedIn StockMarketIndex)
ObjectProperty(hasRate)
  ObjectPropertyDomain(hasRate Security)
  ObjectPropertyRange(hasRate StockMarketRate)
ObjectProperty(hasMetric)
  ObjectPropertyDomain(hasMetric Security)
  ObjectPropertyRange(hasMetric Metric)
ObjectProperty(correlatesWith)
  ObjectPropertyDomain(correlatesWith Security)
  ObjectPropertyRange(correlatesWith Security)

ObjectProperty(hasMetric)
  ObjectPropertyDomain(hasMetric Security)
  ObjectPropertyRange(hasMetric Metric)
  ObjectPropertyDomain(hasMetric Portfolio)

Class(StockMarketRate)
ObjectProperty(onDate)
```

```
  FunctionalObjectProperty(onDate)
  ObjectPropertyDomain(onDate StockMarketRate)
  ObjectPropertyRange(onDate Date)

Class(Metric)
ObjectProperty(isMetricOfSecurity)
Class(Correlation)
  SubClassOf(Correlation Metric)
  SubClassOf(Correlation ObjectAllValuesFrom
    (isMetricOfSecurity Security))
Class(FinancialMetric)
  SubClassOf(FinancialMetric Metric)
Class(SocialMetric)
  SubClassOf(SocialMetric Metric)
```

Онтология[1] определяет следующие основные понятия:

- Portfolio – портфель, составленный из ценных бумаг определённого списка компаний, имеющий, с ним также могут быть связаны метрики оценки портфеля;
- Security – ценные бумаги компании, участвующие в фондовом рынке, у них есть идентификаторы, они могут принадлежать списку фондового индекса, оцениваются котировками, метриками надёжности, могут иметь зависимость от других бумаг;
- StockMarketRate – котировка бумаги, зависящая от времени;
- Metric – метрика для оценки надёжности ценной бумаги или портфеля; одной из метрик оценки надёжности бумаги является корреляция её котировки с другими бумагами.

Особо отметим, что представленная онтология определяет понятия и связи предметной области фондового рынка в отличие от спецификации концептуальной схемы (названной в [6] онтологией области приложения), определяющей представление данных при решении задачи в потоке работ на правилах, хотя и онтология, и концептуальная схема используют выразительные средства, определяемые языком OWL 2. Описания концептуальной схемы недостаточны для использования в метаданных о предметной областью, так как многие понятия отношения предметной области сведены в ней к примитивным типам данных. Подробнее различия и связи онтологий и концептуальных схем предметных областей обсуждаются в [12].

Одновременно с онтологией предметной области для определения метаданных потоков работ используются другие онтологии, определяющие различные аспекты описываемых элементов потоков работ. В частности, для связывания правил с видами элементов потоков работ, которые определены этими правилами, используется онтология структуры потоков работ[2].

```
Class(Workflow)
ObjectProperty(hasTask)
  ObjectPropertyDomain(hasTask Workflow)
  ObjectPropertyRange(hasTask Task)

Class(Task)
ObjectProperty(hasParameter)
  ObjectPropertyDomain(hasParameter Task)
  ObjectPropertyRange(hasParameter TaskParameter)
  InverseObjectProperties(isParameterOf
    hasParameter)
ObjectProperty(hasInputParameter)
  SubObjectPropertyOf(hasInputParameter
    hasParameter)
  ObjectPropertyDomain(hasInputParameter Task)
  ObjectPropertyRange(hasInputParameter
    InputParameter)
  InverseObjectProperties(isInputParameterOf
    hasInputParameter)
ObjectProperty(hasOutputParameter)
  SubObjectPropertyOf(hasOutputParameter
    hasParameter)
  ObjectPropertyDomain(hasOutputParameter Task)
  ObjectPropertyRange(hasOutputParameter
    OutputParameter)
  InverseObjectProperties(isOutputParameterOf
    hasOutputParameter)
```

В приведённом фрагменте онтологии структуры потока работ определены понятия:

- Workflow – поток работ в целом, состоящий из набора деятельностей;
- Task – деятельность, которая может иметь входные и выходные параметры.

Помимо этого онтология определяет разновидности деятельностей, такие как начало и завершение потока, вызов подпотока, шаблоны управления потоками и другие понятия.

В терминах двух представленных онтологий приведём пример аннотации, определяющей метаданные для выходного параметра деятельсности (спецификация представлена в формате RIF XML):

```
<declare><Var>
  <id>
    <Const>GetPortfolios_Output</Const>
  </id>
  <meta>
  <Frame>
    <object>
      <Const>GetPortfolios_Output</Const>
    </object>
    <slot>
      <Const>rdf:type</Const>
      <Const>wf:OutputParameter</Const>
    </slot>
    <slot>
      <Const>wf:isOutputParameterOf</Const>
      <Const>:GetPortfolios</Const>
    </slot>
    <slot>
      <Const>rdf:type</Const>
      <Const>pont:Portfolio</Const>
    </slot>
  </Frame>
  </meta>
?p
</Var></declare>
```

Данная спецификация метаданных определена для переменной `?p` в правиле, соответствующем деятельности потока работ. В первую очередь, она определяет в текущем пространстве имён уникальный идентификатор данного элемента правила RIF (`GetPortfolios_Output`). С этим идентификатором связываются метаданные в

[1] http://ontology.ipi.ac.ru/ontologies/stockmarket.owl
[2] http://ontology.ipi.ac.ru/ontologies/wf.owl

терминах двух определённых выше онтологий (пространство имён `pont` соответствует онтологии предметной области, а `wf` – онтологии структуры потоков работ). Во-первых, определяется, что элемент с данным идентификатором является выходным параметром (экземпляром класса `OutputParameter`) деятельности, решающей подзадачу поиска портфелей (отношение `isOutputParameterOf` к объекту с идентификатором `GetPortfolios`), а также является экземпляром класса `Portfolio`, то есть возвращаемые деятельностью данные должны являться портфелями. Идентификатор `GetPortfolios`, должен быть определён подобным образом в метаданных, связанных с правилом в целом.

Таким образом, метаописание позволяет связать спецификации правил с предметной областью, в которой решается задача, определить части правил, которые соответствуют элементам потоков работ, а также семантически связать элементы друг с другом с помощью выражений в терминах онтологий.

3 Обзор методов, связанных с повторным использованием потоков работ

В большинстве исследований, посвящённых метаданным потоков работ, состав метаданных ограничивается набором предопределённых свойств для работы с простыми сопроводительными данными: именами, вербальными определениями, информацией об авторах, версиях, правах, дате создания и других достаточно ограниченных описаниями [11]. Такие подходы к спецификации метаданных представляются недостаточными для выразительного семантического описания и поиска потоков работ.

Как аннотирование потоков работ метаданными использует простые поля описаний, так же большинство проектов, работающих с потоками работ, ограничиваются методами поиска на основе ключевых слов, относящихся к потокам работ как цельным объектам [5]. Проект wf4ever [10] предоставляет набор средств для поддержки повторного использования, включая аннотирование потоков работ в целом и их компонентов, учитывает в сопровождающих спецификациях происхождение данных, являющихся результатами работы процессов, многоверсионность и другие аспекты. Проект OPM [7] использует развитую модель происхождения данных для выражения семантики воспроизводимости результатов, в том числе, для потоков работ.

В контексте настоящего исследования необходимо упомянуть подходы process mining [4], специализирующиеся, главным образом, на анализе лог-файлов. В исследованиях используются модели процессов, являющиеся спецификациями структуры потоков работ. Записи логов исполняемых деятельностей или происходящих событий сопоставляются моделям процессов. На основе лог-файлов решаются следующие виды задач.

- Под задачей обнаружения потоков работ понимается восстановление фактической структуры потока работ по лог-файлам работы его экземпляра. Таким образом, могут быть вскрыты потоки работ, не имеющие формальных спецификаций модели процесса.
- Задача установления конформности (conformance) потока работ заключается в проверке соответствия модели потока работ данным, получаемым из лог-файлов о работе его реализации.
- Задача усовершенствования модели потока работ отличается от задачи установления конформности тем, что модель не только оценивается на соответствие реальным событиям, но и меняется для более точного соответствия.

Эти исследования рождают множество публикаций с развитием и применением представленных задач. Они полезны для решения задач поиска потоков работ по спецификации их структуры, для описания и дальнейшего повторного использования доступных потоков работ, не имеющих формальной спецификации, но генерирующих лог-файлы во время своей работы, для контроля соответствия реализованных и найденных потоков работ спецификациям.

4 Организация поиска релевантных потоков работ по сформулированным требованиям

Благодаря тому, что в используемой в данном исследовании модели потоков работ спецификации правил, выражающие семантику их поведения, независимы от сопровождающих их метаданных, правила и метаданные могут обрабатываться независимыми инструментами. Спецификации правил используются для реализации потоков работ в определённых системах, исполняющих их в соответствии с семантикой используемых диалектов. Для предварительного связывания элементов спецификаций потоков работ должны использоваться метаданные. Для этого реализуется независимая от спецификаций правил возможность поиска потоков работ по метаданным.

Спецификации фреймов, содержащие значения метаданных, преобразуются в триплеты RDF в соответствии с рекомендациями W3C [3], сохраняются в отдельном хранилище RDF и в дальнейшем используются для запросов поиска по метаданным. В частности фрейм RIF с метаданными, соответствующий XML-представлению в приведённом выше примере:

```
GetPortfolios_Output
  [ rdf:type -> wf:OutputParameter,
    wf:isOutputParameterOf -> GetPortfolios,
    rdf:type -> pont:Portfolio ],
```

будет преобразован в триплеты RDF

```
pwf:GetPortfolios_Output
   rdf:type wf:OutputParameter;
   wf:isOutputParameterOf pwf:GetPortfolios;
   rdf:type pont:Portfolio.
```

Таким образом, база триплетов собирает в себе набор метаданных и идентификаторов, по которым можно установить, с какими именно элементами спецификации потоков работ на правилах связаны определённые метаданные. В качестве RDF-словарей может использоваться произвольный набор онтологий, в частности определяющих состав метаданных, разработанный в [13]. Поиск потоков работ и их фрагментов по метаданным организуется с помощью задания запросов на языке SPARQL [8] к базе триплетов, содержащей метаданные.

Запросы на языке SPARQL формулируются в соответствии с требованиями задачи, которая должна быть решена в предметной области, либо с требованиями спецификации потока работ, который необходимо реализовать с помощью повторного использования существующих потоков работ, их фрагментов и доступных сервисов.

5 Поиск релевантных потоков работ в целом и их фрагментов

Поиск потоков работ для обеспечения их повторного использования при наличии метаданных, требуемых в [13], производится на основании соответствия выбранных или всех одновременно критериев:

- соответствие потока работ понятиям или выражениям в терминах понятий онтологии предметной области, описывающих зависимости/функции, методы, процессы, могущие применяться в данной предметной области;
- соответствие понятий или выражений в терминах понятий, описывающих входные и/или выходные параметры потоков работ (например, для поиска методов, которые из определённого набора параметров получают требуемый тип результата);
- выполнение требований к качеству входных данных и качеству возвращаемых результатов потока работ в терминах онтологии качества данных (например, требования актуальности);
- требования к происхождению потока работ (например, по автору разработанных реализаций);
- требования к происхождению входных данных (например, определённое оборудование, которым собраны первичные данные наблюдений).

Таким образом, требования к искомым в коллекции научных методов потокам работ могут затрагивать как функциональность реализуемых ими научных методов, так и предусловия и постусловия, выраженные в терминах онтологий, а также требования к надёжности применяемых методов, используемых данных и получаемых результатов.

Информация о происхождении и качестве данных и методов в потоках работ используется для

- спецификации достоверности, полноты, точности требуемых данных и достигаемых результатов
- контроля реальных источников данных и их качества в соответствии с требованиями задачи;
- контроля соответствия требованиям решения задачи используемых открытых реализаций научных методов.

Помимо этого, решение научных задач предметной области может выбираться как фрагментарно из других потоков работ, так и из отдельных фрагментов потоков работ и из существующих сервисов. Необходимый фрагмент обработки данных может оказаться частью реализации потока работ, решающего в целом отличную задачу. Для этого требования в запросах формулируются не к потокам работ в целом, а к параметрам деятельностей в составе потоков работ.

В качестве примера зададим запрос для поиска потоков работ, реализующих метрики оценки надёжности портфелей ценных бумаг.

```
select distinct ?task1 ?task2 where
{
  ?task1 rdf:type pont:Metric .
  ?in1 wf:isInputParameterOf ?task1 .
  ?in1 rdf:type pont:Security .
  ?var wf:isOutputParameterOf ?task1 .
  ?var rdf:type pont:Metric .
  ?in1 pont:hasMetric ?var .
  ?task2 rdf:type pont:Metric .
  ?var wf:isInputParameterOf ?task2
  ?in2 wf:isInputParameterOf ?task2 .
  ?in2 rdf:type pont:Portfolio .
  ?in2 pont:includesSecurity ?in1 .
  ?out2 wf:isOutputParameterOf ?task2 .
  ?out2 rdf:type pont:Metric .
  ?in2 pont:hasMetric ?out2 .
}
```

По условию запроса необходимо найти деятельности, одна из которых принимает на вход объекты ценных бумаг, вычисляет и возвращает для него некоторую метрику, а вторая деятельность принимает на вход результаты первой деятельности, и вычисляет обобщающую метрику для портфеля, содержащего ценные бумаги, для которых вычислена первая метрика.

Представим, что в базе триплетов хранятся метаданные следующих деятельностей:

- getPositiveTweetRatio – вычисляет тональность сообщений о ценной бумаге в Twitter;
- computePortfolioTwitterMetrics – на основе тональности сообщений о ценных бумагах вычисляет тональность отношения к содержащему их портфелю;
- getSecurityFinancialMetrics – вычисляет метрику надёжности ценной бумаги, учитывающую выгоду и риски на основе истории котировок;
- computePortfolioFinancialMetrics – для портфеля в целом, содержащего ценные бумаги, вычисляет обобщённую финансовую метрику.

Например, для одной из деятельностей и её структурных элементов хранятся следующие триплеты:

```
fin:getSecurityFinancialMetrics
   rdf:type wf:Task;
   rdf:type pont:FinancialMetric;
   wf:hasInputParameter fin:finMetricPar;
   wf:hasOutputParameter fin:securityPar;
fin:securityPar
   rdf:type pont:Security;
   rdf:type wf:InputParameter;
fin:finMetric
   rdf:type pont:FinancialMetric;
   rdf:type wf:OutputParameter;
```

При условии адекватного описания метаданными спецификаций деятельностей и их связей друг с другом внутри потоков работ ответ на запрос будет содержать следующие кортежи:

```
<sparql xmlns=http://www.w3.org/2005/sparql-results#>
   <head>
      <variable name="task1"/>
      <variable name="task2"/>
   </head>
<results>
   <result>
      <binding name="task1">

   <uri>http://ontology.ipi.ac.ru/portfolio.rif#
        getPositiveTweetRatio</uri>
      </binding>
      <binding name="task2">

   <uri>http://ontology.ipi.ac.ru/portfolio.rif#
        computePortfolioTwitterMetrics</uri>
      </binding>
   </result>
   <result>
      <binding name="task1">

   <uri>http://ontology.ipi.ac.ru/portfolio.rif#
        getSecurityFinancialMetrics</uri>
      </binding>
      <binding name="task2">

   <uri>http://ontology.ipi.ac.ru/portfolio.rif#
        computePortfolioFinancialMetrics</uri>
      </binding>
   </result>
</results>
</sparql>
```

Таким образом, найдены спецификации деятельностей, которые можно использовать повторно для реализации метрик ценных бумаг и портфелей при решении задачи выбора наилучшего портфеля.

Спецификация потока работ, использующего найденные спецификации, может быть следующей (рис. 1) [6]:

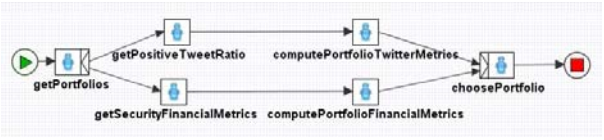

Рис 1. Поток работ для решения задачи выбора лучшего портфеля ценных бумаг

В общем случае выделение релевантных фрагментов является нетривиальной задачей, требующей решения в соответствии со структурой и семантикой выполняемых действий каждым компонентом в составе фрагмента. Эта задача привлекает и метаданные, и сравнение спецификаций шаблонов, и проверку конформности спецификаций и реализаций, и работу экспертов.

Поддержание при спецификациях потоков работ на правилах стиля, при котором управляющая часть потока работ набирается не из произвольных правил, а из наборов правил, реализующих определённые известные шаблоны [9], может упростить проверку конформности. Целесообразно ввести также метаданные, обозначающие те или иные шаблоны в терминах онтологии структуры потоков работ.

6 Семантический контроль используемых методов и принятых решений

Метаданные целесообразно использовать не только при поиске потоков работ и их компонентов в коллекциях научных методов, но и для дальнейшей проверки совместимости семантики данных и интероперабельности потоков работ и фрагментов при объединении найденных и выбранных компонентов для реализации решения научных задач.

Для этого необходимо проводить следующие проверки:

- корректность включения в качестве деятельности данного потока работ существующих компонентов в качестве подпроцессов по их входным и выходным параметрам;
- соответствие семантики входных компонентов семантике входных данных и соответствие выходных данных выходным параметрам по понятиям предметной области, требованиям к качеству, происхождению и другим возможным критериям, учтённым с помощью онтологий;
- соответствие семантики данных, проходящих из выхода одного компонента на вход другого.

Эти проверки выполняются по принципу спецификаций пред- и постусловий: постусловие выхода предыдущего компонента должно быть строже предусловия входа последующего компонента. Требования могут включать как выражения в терминах понятий предметной области, так требования качества и происхождения данных.

7 Проведение экспериментов и проверка интероперабельности потоков работ на тестовых наборах данных

Требования к релевантности по метаданным могут быть в той или иной степени выразительными, а структурное соответствие само по себе не включает проверку семантики компонентов. К тому же реализации спецификаций могут использовать разные инструменты, и по деталям работы они могут отличаться друг от друга. Поэтому для надёжного повторного использования

реализаций потоков работ необходимо проверять их на определённых наборах тестовых данных.

Тесты включают набор данных и требования к ожидаемым результатам, достаточные для проверки всех возможных особых случаев, могущих возникать в потоке работ. Помимо входных и проверочных выходных данных тесты могут включать метаданные.

При тестировании производится контроль прохождения тестов по определённому пути в потоке работ в зависимости от входных данных. Для этого в состав тестов включаются метаданные происхождения данных. Метаданные происхождения, сгенерированные в результате прогона потока работ на тестовых данных, проверяются на соответствие происхождения данных в составе тестов.

Проверяется соответствие результатов требованиям качества, предоставляемым в спецификациях тестов или специфицирующих выходные параметры потока работ.

Помимо этого, требования тестов могут налагаться и на описания исполняемых сред. Для этого также используются метаданные на основе онтологий описания исполняемых сред [13].

Другой подход тестирования реализаций потоков работ, собранных на основе спецификаций, предполагает генерацию лог-файлов при прохождении тестов, активизирующих все возможные пути в потоке работ. При реализации правил RIF компонентами, использующими различные системы вывода, лог-файлы должны генерироваться каждой из них. Решение задачи установления конформности [4] спецификации потока работ и получившейся реализации позволяет подтвердить их соответствие друг другу.

8 Заключение

Работа посвящена организации семантического поиска потоков работ и их фрагментов по метаданным с целью их повторного использования. Она является продолжением исследования, представленного в [13], применяемого к другим техническим условиям. В качестве модели потоков работ [6] используются языки на правилах, что даёт богатые возможности в повышении выразительности спецификаций и в применимых методах анализа потоков работ. В статье разработан подход к представлению и обработке метаданных в данной модели потоков работ. Упор делается на сценариях применения метаданных потоков работ для поиска потоков работ с целью их повторного использования и для проверки их релевантности и интероперабельности.

Благодарности

Работа выполнена при поддержке грантов РФФИ 13-07-00579, 14-07-00548 и Программы Президиума РАН.

Литература

[1] OWL 2 Web Ontology Language Document Overview (Second Edition) – W3C, 2011. – URL: http://www.w3.org/TR/owl-overview/

[2] RIF Overview. – W3C, 2013. – URL: http://www.w3.org/TR/rif-overview/

[3] RIF RDF and OWL Compatibility. – W3C, 2013. – URL: http://www.w3.org/TR/rif-rdf-owl/

[4] W.M.P. Van der Aalst. Process mining: Discovery, Conformance and Enhancement of Business Processes. Springer, Heidelberg, 2011.

[5] C.A. Goble, D.C. De Roure. 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.

[6] L. Kalinichenko, S. Stupnikov, A. Vovchenko, D. Kovalev. Multi-dialect Workflows // ADBIS’2014. – 2014. – LNCS 8716. – P. 352–365.

[7] L. Moreau. Provenance-based reproducibility in the semantic web // Web semantics: science, services and agents on the World Wide Web. – 2011. – Vol. 9, No. 2. – P. 202–221.

[8] Polleres A. SPARQL1. 1: New features and friends (OWL2, RIF) // Web Reasoning and Rule Systems. – Springer Berlin Heidelberg, 2010. – С. 23–26.

[9] N. Russell, A.H.M. ter Hofstede, W.M.P. van der Aalst, and N. Mulyar. Workflow Control-Flow Patterns: A Revised View. – BPM Center Report BPM-06-22, BPMcenter.org. – 2006.

[10] S. Sanchez, et al. WF4Ever: Supporting for reuse and reproducibility in experimental science // EGI Technical Forum. – 2012.

[11] C. Tejo-Alonso et al. Metadata for web ontologies and rules: Current practices and perspectives // Metadata and Semantic Research. – Springer Berlin Heidelberg, 2011. – С. 56–67.

[12] А.Е. Вовченко и др. От спецификаций требований к концептуальной схеме // RCDL'2010. – Казань: КФУ, 2010. – С. 375–381.

[13] Н.А. Скворцов, Д.О. Брюхов, Л.А. Калиниченко, Д. Ковалёв, С.А. Ступников. Метаданные о научных методах для обеспечения их повторного использования и воспроизводимости результатов // RCDL’2013. – Ярославль, 2013.

An Approach to Search of Workflows by Metadata

Nikolay A. Skvortsov

The work is dedicated to methods of search of workflow implementations and their components for reuse by metadata specifications. Workflow specifications are formulated in the RIF language dialects, metadata is represented as RIF annotations. A set of metadata needed for workflow reuse is applied in various tasks during workflow development. The paper demonstrates methods of metadata specifications and semantic search of workflows using them.