Методология организации решения задач над множественными распределенными неоднородными источниками информации[1]

Л.А.Калиниченко, ИПИ РАН, leonidk@synth.ipi.ac.ru

1. Введение

В различных областях науки наблюдается экспоненциальный рост объема получаемых экспериментальных (наблюдательных) данных. Например, в астрономии текущий и ожидаемый темп роста данных от обсерваторий удваивается в течение периода от шести месяцев до одного года. Это более быстрый темп, чем увеличение производительности компьютерных чипов, удваиваемой (согласно закону Мура) каждые 18 месяцев. Сложность использования таких данных увеличивается еще и вследствие их естественной разнородности. Число организаций, получающих данные наблюдений в отдельных областях науки в мире, велико. Разнообразие (информационная несогласованность) получаемой информации вызывается, в частности, не только большим числом организаций, производящих наблюдения, и их независимостью, но и разнообразием объектов наблюдения и непрерывным и быстрым совершенствованием техники наблюдений, вызывающим адекватные изменения структуры и содержания накапливаемой информации.

Чрезвычайно быстро развивается также программный инструментарий, включающий многообразные сервисы для поддержки различных видов обработки информации при решении научных задач и проведении исследований. Такие сервисы производятся различными научными организациями, их описания неоднородны и неполны.

Увеличивающийся разрыв между исследователями и источниками данных и сервисов приводит к необходимости поиска новых путей организации решения задач над множественными распределенными коллекциями данных и программ, которые концентрируются в специализированных центрах данных и вычислительных ресурсов. Разработан (разрабатывается) ряд инфраструктур, которые технически позволяют реализовать решение задач над множественными информационными источниками. Среди них Веб сервисы, Грид-архитектуры, Семантический Веб, технологии распределенных баз данных, интероперабельные технологии промежуточного слоя, и др. Они составляют техническую среду для организации решения задач.

Традиционно при решении задач специалисты используют привычные для них источники информации, и формулируют задачи, учитывая лишь такие источники. Подобные способы формулирования и решения задач называются далее *движимыми* конкретными *источниками* информации, отобранными до или, в лучшем случае, в процессе формулирования задачи. Очевидна неполнота информации, которую удается охватить при таком подходе. Множество источников данных и сервисов, существующих в Интернете, их разнообразие, вызывают потребность в радикальном изменении такого традиционного подхода. Существо этого изменения заключается в том, что задачи должны формулироваться независимо от существующих источников информации, и лишь после такой формулировки, должна осуществляться идентификация релевантных задаче источников, приведение их к виду, требуемому в задаче, их интеграция, идентификация

[1] Настоящая работа выполнена при поддержке РФФИ, грант № 05-07-90413-в, и Программы "Фундаментальные основы информационных технологий и систем" ОИТВС РАН

сервисов, которые позволяют реализовать отдельные части абстрактного процесса решения задачи. Только после этого должно осуществляться конструирование конкретного процесса решения на основе отобранных источников данных и сервисов.

Этот подход, называемый *ориентированным на проблему*, должен значительно больше опираться на использование техники представления знаний, нежели традиционный. Так, формулирование задачи должно быть основано на определении ее проблемной области, включающем ее терминологию и систему понятий, абстрактное описание соответствующей материальной системы, определение адекватных моделей и теорий, абстрактное описание требуемых в задаче характеристик объектов реального мира, определение методов, алгоритмов и процессов решения задачи.

Одной из проблем при таком подходе остается та, что во всех названных выше инфраструктурах до сих пор открытым является вопрос интегрированного представления множественных источников информации для исследователя, решающего задачу. Здесь также существуют два принципиально разных решения: двигаясь от источников к задачам (создается интегрированное представление множества источников независимо от задач) и от задачи к источникам (создается описание предметной области класса задач, в которое отображаются релевантные задаче источники информации). При первом подходе трудно обеспечить масштабируемость по числу источников. Например, в астрономии число источников (архивов, каталогов) достигает многих тысяч. Если применяется интегрированная схема совокупности источников, ее приходится изменять при включении в рассмотрение каждого нового источника. Определение мультибазы данных (каждому источнику в глобальной схеме соответствует своя подсхема), глобальная схема становится необозримой для исследователя.

Другой подход предусматривает создание *предметных посредников*, поддерживающих взаимодействие между исследователем и источниками посредством описания предметной области класса задач (в терминах понятий, структур данных, функций и процессов решения задач). При этом предполагается, что информационные источники опубликованы в коллективных хранилищах, а операции идентификации нужных источников являются их основными операциями. Для этого в коллективных хранилищах поддерживаются метаданные для описания хранимых в них информационных источников.

Методы организации решения задач на основе подхода, ориентированного на проблему, при использовании техники предметных посредников кратко рассматриваются в настоящей статье.

2. Основания и инструментарий подхода к решению задач, ориентированного на проблему

2.1. Композиции компонентов на основе методов и инструментов теории уточнения

Проектирование информационной системы (ИС) для решения задач над множественными неоднородными источниками информации является *композиционным* [2], основная идея которого состоит в том, чтобы построить композицию спецификаций существующих, релевантных задаче компонентов (информационных, программных, процессных), так, чтобы она уточняла более абстрактную спецификацию разрабатываемой ИС. В целях проектирования, спецификации компонентов и ИС приводятся к однородному представлению в *канонической информационной модели*. Принципиальной составляющей процесса композиционного проектирования является формальное доказательство факта

уточнения спецификации ИС композицией спецификаций компонентов [1]. Уточнение системой *B* системы *A* означает, что пользователь может использовать систему *B* вместо системы *A*, не замечая факта замены *A* на *B*. Формальное доказательство уточнения позволяет утверждать, что сконструированная из существующих компонентов система действительно реализует абстрактную спецификацию ИС. Для обеспечения процесса доказательства уточнения в Лаборатории композиционных методов проектирования информационных систем ИПИ РАН (далее для краткости ЛКМП ИПИ РАН) разработаны [18]:

- формальная семантика канонической информационной модели (языка СИНТЕЗ [8]) в Нотации Абстрактных Машин (Abstract Machine Notation, AMN [1]). В качестве ядра канонической информационной модели, предназначенной для унифицированного представления спецификаций ИС и спецификаций компонентов при композиционном проектировании систем, используется гибридный объектный язык, включающий средства спецификации как структурированных, так и слабоструктурированных данных. AMN представляет собой формальный язык спецификаций, основанный на логике предикатов первого порядка и теории множеств. AMN предназначена для построения математических моделей ИС. AMN поддерживается специальной технологией (B-technology), включающей инструментальные средства (B- Toolkit, Antelier B) формализации и автоматизированного доказательства корректности уточнения, успешно используемые в ряде индустриальных проектов,

- инструментальные средства, осуществляющие автоматическое отображение спецификаций канонической модели в AMN.

Благодаря определению формальной семантики канонической модели (языка СИНТЕЗ) достигнута возможность совместного доказательства уточнения структурных, функциональных и процессных свойств спецификации типов ИС структурными, функциональными и процессными свойствами спецификации типов существующих информационных источников (или композиции спецификаций таких типов) [2].

2.2. Синтез канонических информационных моделей

Настоящий период развития информационных технологий (ИТ) характеризуется взрывоподобным процессом создания разнообразных моделей представления информации. Это развитие происходит как в рамках конкретных распределенных инфраструктур (таких как архитектуры OMG (в частности, архитектуры, движимые моделями представления информации (MDA)), архитектуры семантического Web и Web сервисов, архитектуры электронных библиотек как коллективных хранилищ информации в различных предметных областях, архитектуры информационных грид), так и в стандартах языков и моделей данных (таких как, например, ODMG, SQL, UML, стеки XML и RDF моделей данных), процессных моделей и моделей потоков работ, семантических моделей (включая онтологические модели и модели метаданных), моделей цифровых репозиториев данных и знаний в конкретных областях науки (например, виртуальные обсерватории в астрономии). Этот процесс сопровождается другой тенденцией – накоплением использующих подобные модели источников данных и сервисов, число которых экспоненциально растет. Этот рост вызывает все увеличивающуюся потребность интеграции модельно неоднородных информационных источников в различных применениях, а также их повторного использования и композиции для реализации новых информационных систем. Указанные тенденции противоречивы: чем больше разнообразие применяемых моделей в различных компонентах и сервисах, тем более сложными становятся проблемы их интеграции и

композиции. Эти тенденции не новы, но с течением времени разнообразие различных моделей и их сложность растет вместе с ростом потребности достижения интеграции и композиции разномодельных информационных источников. Масштабы этих явлений, определяющих возможности конструирования распределенных информационных систем в различных областях, повторного использования, трейдинга и композиции компонентов, достижения их семантической интероперабельности (совместной работы в конкретных применениях), интеграции неоднородных информационных источников, являются достаточной мотивацией для исследования и разработки адекватных методов оперирования разнообразными моделями представления информации. Основу этих методов составляет понятие *канонической информационной модели*, служащей в качестве общего языка, «эсперанто», для адекватного выражения семантики разнообразных информационных моделей, окружающих нас. Для доказательства того, что определение в одном языке может быть заменено на определение в другом, предоставляются средства формальной спецификации и коммутативные отображения моделей. Исторически сначала развивались идеи отображения моделей данных и построения канонической модели для структурированных моделей данных. Были введены основополагающие определения эквивалентности состояний баз данных, схем баз данных и моделей данных для того, чтобы при построении отображений разнообразных структурированных моделей данных в каноническую сохранялись операции без потери информации [6,7]. Каждая модель данных при этом определялась синтаксисом и семантикой двух языков – языка определения данных (ЯОД) и языка манипулирования данными (ЯМД). Основным принципом отображения произвольной исходной модели данных в целевую модель (каноническую) явился *принцип коммутативного отображения моделей данных*, согласно которому сохранение операций и информации исходной модели данных при ее отображении в каноническую достигается при условии, что диаграмма отображения ЯОД (схем) и диаграмма отображения ЯМД (операторов) являются коммутативными [6]. При этом в процессе конструирования отображений моделей данных в качестве формализма (метамодели) использовалась денотационная семантика, позволявшая доказывать коммутативность указанных диаграмм [6]. Такое доказательство приходилось проводить вручную.

Позднее, для объектных моделей данных, метод отображения моделей данных и построения канонических моделей был видоизменен следующим образом. В качестве формализма (метамодели) метода вместо денотационной семантики была применена Нотация Абстрактных Машин (AMN), позволяющая определять теоретико-модельные спецификации в логике первого порядка и осуществлять доказательство факта уточнения спецификаций [1]. Теория уточнений позволила развить основополагающие определения отношений между типами данных, схемами данных, моделями данных так, чтобы вместо эквивалентности соответствующих спецификаций, можно было рассуждать об их уточнении [9]. Наличие специальных инструментов для AMN (B-технология) позволяет осуществлять доказательство коммутативности отображений интерактивно: необходимые для доказательства уточнений теоремы генерируются B автоматически, а их доказательство (в общем случае) реализуется с помощью человека. Основной принцип синтеза канонических моделей состоит в том, что необходима *расширяемая* каноническая модель для семантической интеграции и интероперабельности информации в разнородной среде, включающей различные модели. Ядро канонической модели фиксируется. Для каждой конкретной информационной модели M_i среды определяется расширение ядра канонической модели, так, что оно вместе с ядром *уточняется* M_i . Такая уточняющая трансформация моделей должна быть *доказуемо правильной*. Каноническая модель среды синтезируется как *объединение расширений* , образованных для моделей M_i среды. Этот подход был применен недавно также для синтеза канонической модели процессов, охватывающей известные модели потоков работ [13].

Использование канонической модели с формальной семантикой, в которой возможно проведение полного доказательства факта уточнения, позволяет систематически рассмотреть проблему организации решения задач в среде множественных распределенных неоднородных источников информации и разработать базовые методы и средства, рассматриваемые далее.

2.3. Идентификация релевантных источников информации (данных и программных сервисов) и их регистрация в посреднике

Известные попытки решения задачи идентификации релевантных спецификации ИС источников информации заключаются в следующем.

Исследования в области решеток типов и соответствующих алгебр имеют весьма продолжительную историю. Как правило, всегда имеется стремление к достижению компромисса между разумной выразительностью спецификаций и разрешимостью. При регистрации в предметном посреднике разрешимость приносится в жертву ради достижения полноты спецификаций. Благодаря полноте спецификаций, достигается хорошо обоснованный способ идентификации общих фрагментов спецификаций типов, обеспечивающей возможность их адекватной композиции и повторного использования. В этой области сравнительно немного работ. В Австрии (в Клагенфуртском университете) предложено исчисление структур данных, основанное на упорядочении множества спецификаций типов на базе отношения поглощения (subsumption) спецификаций и формировании соответствующей решетки [16]. На этой основе рассмотрена структура репозитория спецификаций компонентов как информационно-поисковой системы. Проблема поиска компонентов рассматривается узко - для компонентов-функций, представляемых отношениями, содержащими все допустимые пары входных/выходных значений функций. Порядок на основе отношения уточнения, заданный на множестве функций, имеет свойства решетки. Решетка формируется посредством операций *join* и *meet* на отношениях, представляющих функции. *Join (meet)* представляют суммарную информацию (общую информацию), содержащуюся в таких отношениях.

В MIT предложен способ сопоставления сигнатур операций как механизм поиска спецификаций программных компонентов в их репозитории [21]. Эта работа расширяема на случай представления спецификаций функций их пред- и пост- условиями.

Проблема разрешения структурных конфликтов при идентификации близка проблеме разрешения конфликтов при интеграции схем баз данных. В Пенсильванском университете рассматривались два подхода к разрешению структурных конфликтов [15]: использование предопределенных правил преобразования или языка высокого уровня для описания преобразований. При использовании предопределенных правил преобразования, целью является автоматическое разрешение структурных конфликтов и генерация интеграционной схемы. Разработчик задает соответствие между элементами (классами, атрибутами) схем и соответствие между путями в схемах. Затем на основе предопределенных правил генерируется интеграционная схема (классы и функции преобразования значений между классами локальных и интеграционной схем). Основными преимуществами данного подхода являются простота аргументации и доказательность корректности применения данных правил. В частности, можно доказать, что каждое правило сохраняет информацию, и что комбинация таких правил также сохраняет информацию. Основным недостатком этого подхода является фиксация используемых правил, что накладывает ограничения на допустимые преобразования.

При использовании языка высокого уровня, разработчику предоставляется богатый язык для описания правил преобразования. Однако, каждая функция разрешения конфликта должна быть запрограммирована, и ее правильность доказана отдельно. Как и в предыдущем подходе, разработчик сам задает соответствие между элементами схем. Этот подход предоставляет пользователю более гибкие средства формирования интеграционной схемы, позволяющие разрешать любые конфликты.

Недавняя статья [14] содержит оценку последних работ в области идентификации релевантных запросу сервисов. В ней предлагается использовать теоретико-множественную модель запросов и сервисов и разбить процесс идентификации на два этапа – сначала реализуется приблизительная идентификация кандидатов на основе подобия их сигнатурных описаний, а затем осуществляется использование более полной информации (возможно, на основе дескриптивной логики) для окончательного определения адекватности сервиса. Второй этап является значительно более трудоемким. Оценка других работ (таких как OWL-S, METEOR-S), сводится к тому, что в них предпринимается попытка использования дескриптивной логики и решить задачу одноэтапно, а также отсутствует необходимый уровень концептуального моделирования для полного решения задачи адекватности сервиса.

Настоящая работа особое внимание уделяет именно этой проблеме, и является свободной от названных недостатков. Определение предметного посредника и регистрация в нем информационных источников рассматриваются в ЛКМП ИПИ РАН как задача композиционного проектирования систем [3]. Регистрация источников есть процесс целенаправленной трансформации спецификаций, включающий декомпозицию спецификаций посредника на непротиворечивые фрагменты, поиск среди спецификаций релевантных источников подходящих типов данных - кандидатов для уточнения ими спецификаций типов посредника, построение выражений, определяющих классы источников в виде композиции классов посредника. Для подобного манипулирования спецификациями разработано специальное *исчисление спецификаций* [10]. В нем предложен принцип декомпозиции спецификаций типов в набор *редуктов* спецификаций типов, служащих основными единицами повторного использования и композиции. Определена операция определения *наибольшего общего редукта* спецификаций типов компонентов и требований. На основе частично упорядоченного множества спецификаций типов определены также решетка и алгебра типов. Эти структуры послужили теоретической базой для разработки репозитория метаинформации, хранящего спецификации требований, компонентов, а также промежуточные спецификации, возникающие в процессе композиционного проектирования. Репозиторий метаинформации, представленной на языке СИНТЕЗ, реализован на основе СУБД Oracle.

Принципиальным моментом в этой схеме является реализация доказательства уточнения на основе математической модели спецификаций посредника и источников, а также инструментальных средств, реализующих разработанные алгоритмы отображения моделей спецификаций канонической модели в AMN.

Идентификация релевантных источников (предшествующая регистрации) основана на использовании трех моделей – *модели метаданных*, характеризующих свойства источников информации, собранных в некотором коллективном хранилище, *онтологической модели*, позволяющей формально определять понятия предметной области, и *канонической модели*, позволяющей формально определять структуру и поведение объектов предметной области задачи и информационных источников. Рассуждения в канонической и в онтологической модели основаны на семантике канонической модели и средствах доказательства уточнения. При этом в онтологической

модели необходимо достичь согласования понятийной семантики спецификаций посредника и регистрируемых источников информации. Рассуждения в модели метаданных являются эвристическими на основе нефункциональных требований к требуемым в классе задач источникам (к таковым относятся, в частности, показатели качества данных в источниках). Необходимые модели метаданных и алгоритмы поиска составляют часть метода. В целях проектирования, спецификации посредника и источников задаются в однородном их представлении в канонической модели, хотя для этого может потребоваться преобразование в такую модель из некоторого другого языка спецификаций, например из UML.

Сложной проблемой композиционного проектирования является согласование прикладных контекстов разрабатываемой ИС (соответствующей проблеме) и конкретных источников. В ЛКМП ИПИ РАН такое согласование осуществляется на основе онтологического подхода. Онтологическая спецификация представляет собой множество определений понятий конкретной предметной области в форме, доступной как машине так и человеку. Онтологические спецификации играют роль 'клея' фрагментов компонентов для их семантической композиции. Онтологические определения аннотируют элементы спецификаций посредника и спецификаций источников (заданных в форме типов, классов, процессов). В существующем подходе используется прежде всего вербальная (задаваемая подобно определениям терминов в толковом словаре) форма определения понятий. Онтологические спецификации являются частью упомянутого выше репозитория метаинформации. Вербальное представление онтологии дополняется более формальными спецификациями на основе абстрактных типов данных канонической модели и техники доказательства уточнения.

Перечисленные подходы положены в основу разработанного в ЛКМ ИПИ РАН прототипа средств идентификации и регистрации источников информации в посреднике (на основе алгоритмов композиционного проектирования систем) [3]. При этом онтологические спецификации используются для идентификации классов посредника, семантически релевантных классам источника. Максимальное подмножество информации класса источника, релевантное классу посредника, устанавливается на основании максимального общего фрагмента спецификаций соответствующих типов экземпляров этих классов. Конкретизирующие типы, устраняющие возникающие конфликты (значений, структур данных и поведения) в названных типах экземпляров, определяются так, чтобы тип экземпляра класса посредника уточнялся бы типом экземпляра класса источника. Основным результатом регистрации является выражение, определяющее, как класс источника выражается через посредство классов посредника.

2.4. Методы и средства формулирования и реализации задач (запросов) над множественными источниками информации

Общий подход заключается в формулировании задачи в терминах спецификации предметного посредника и преобразовании этой формулировки во множество задач (запросов) к источникам информации, зарегистрированным в посреднике. Такое преобразование в теории баз данных известно как *переписывание запросов на основе взглядов* (источники информации трактуются как материализованные взгляды над виртуальными классами посредника) [5]. Это сложная задача, имеющая решения применительно к конкретным моделям данных (как правило, с рядом ограничений).

Среди многочисленных исследований алгоритмов обеспечения ответа на запрос над материализованными взглядами, в контексте настоящей работы применяются алгоритмы, использующие *инверсные правила* [5,20]. Такие алгоритмы отличаются концептуальной

простотой, модульностью и способностью порождать максимально-включенные запросы за время полиномиальное по отношению к размерам спецификации запросов и взглядов. Особенно важными являются разработанные недавно алгоритмы переписывания объединений конъюнктивных запросов [20], которые обладают рядом преимуществ по сравнению с известными ранее (такими как MiniCon, U-join, Bucket). Эти алгоритмы потребовали существенного развития для типизированной объектно-ориентированной модели данных.

Метод переписывания запросов в типизированной объектной среде, использующий технику инверсных правил, разработан в ЛКМП ИПИ РАН [12]. Этот метод основан на использовании отношения уточнения между типами данных посредника и типами данных источников, что приводит к включению переписанных запросов в оригинальный запрос задачи (говорят, что запрос *Q* включается в запрос *Q*', если он продуцирует подмножество ответов на *Q*' для любой базы данных).

3. Методология решения задач, ориентированного на проблему

Методология должна включать совокупность принципов, правил и методов описания спецификации предметной области и постановок задач в среде множественных распределенных неоднородных источников информации на основе канонической модели данных.

3.1. Формулирование задачи и спецификация посредника

Проблемная область в естественных науках (Рис. 1) определяется терминологией и понятиями, соотнесенными соответствующей материальной системе, определением наблюдаемых свойств материальных объектов, определениями моделей и теорий, разработанных для проблемной области, интерпретациями теорий в системе наблюдаемых значений характеристик реальных объектов (представленных в многочисленных базах данных), имитационными моделями и их интерпретациями. Спецификации методов решения задач, алгоритмов, программного инструментария обеспечивают возможности применения информационных технологий при решении конкретных задач.

Спецификация ИС для решения конкретной задачи (или, точнее, класса задач) включает определения специфической терминологии и понятий в данном классе, выражаемых соответствующими словарями и онтологическими определениями в выбранных для этого специальных языках (например, Ontolingua или OWL), спецификации классов объектов, соответствующих предметной области (задаче), спецификации типов экземпляров названных классов и их методов, определяющих поведение подобных классов, спецификации процессов решения задач данного класса как совмещенных во времени последовательностей действий, реализуемых методами классов. Спецификация классов, их методов и процессов может осуществляться средствами произвольных языков спецификаций (например, UML). Однако, предполагается, что затем такие спецификации (включая онтологические) преобразуются в спецификации канонической модели, имеющей формальную семантику. Следует заметить, что формулирование ИС для решения задачи производится в ориентации на проблему вне зависимости от конкретных существующих источников информации. Результат этой деятельности, выполняемой заинтересованным научным сообществом, составляет спецификацию посредника, образуемую в результате достижения консенсуса в таком сообществе, а сама деятельность по спецификации посредника называется периодом его *консолидации*.

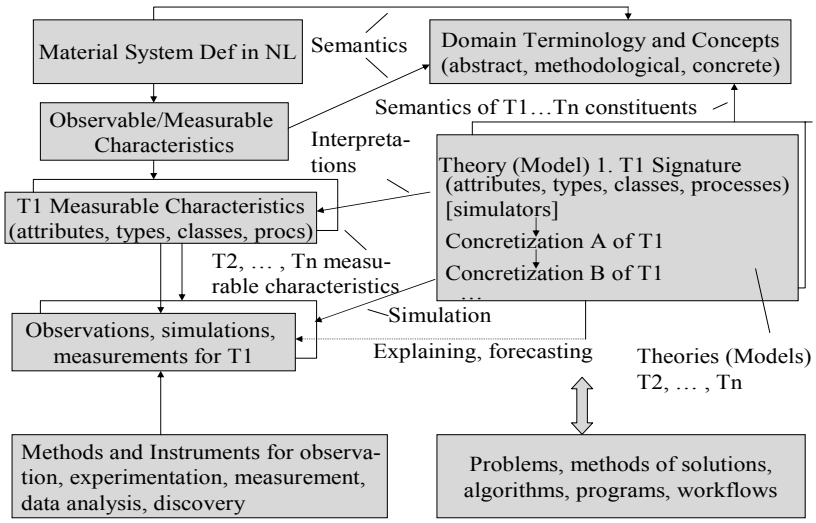

Рис. 1. Спецификация проблемной области в естественных науках

3.2. Поиск информационных источников на основе метаданных

Спецификация посредника определяет понятия, структуру информации и поведение, необходимые для решения задач посредника. Наряду с этим при определении посредника необходимо также задание *модели требований* к необходимой информации. Модель требований задается в рамках той же канонической модели и онтологии посредника. Требования к необходимой информации задаются в виде предикатов над схемой посредника и схемой требований. В модели требований, в частности, добавляются нефункциональные требования к информационным источникам для класса задач (например, требуемые характеристики качества данных, такие как точность). Для спецификации метаданных модели требований вводятся *типы метаданных*. Тип метаданных, определенный для посредника, является частью его модели требований и может не совпадать с типами метаданных потенциальных источников. Сами требования вводятся как *инварианты* типа метаданных. Действие модели требований на классы посредников распространяется посредством *метаклассов*, экземплярами которых становятся классы посредника. Каждый класс посредника может быть экземпляром одного или нескольких метаклассов.

Релевантность метаданных источника требованиям посредника определяется посредством уточнения соответствующих типов: редукт типа метаданных источника должен уточнять редукт соответствующего типа метаданных посредника, играющего роль требований к информации, заданных в посреднике.

Предполагается, что все источники изначально зарегистрированы в *реестрах* коллективных хранилищ в Интернете. В каждом реестре источник представляется идентификатором, своими метаданными, семантика которых описывается определенной онтологией, и другой дополнительной информацией. Такие реестры не зависят от существования посредников. В каждом реестре осуществляется поиск источников – *кандидатов* на регистрацию в посреднике. Этот этап представляет собой предварительный поиск, необходимый для сужения множества рассматриваемых далее источников

3.3. Фильтрация отобранных источников на основе онтологических определений

Основой для онтологически семантического поиска классов и типов спецификаций информационных источников, подходящих для их регистрации в посреднике, является онтологическая модель [17,11]. *Онтологическое понятие* отражает существенные свойства, связи и отношения класса объектов реального мира, воспринимаемые агентами в данной предметной области. Для определения онтологии конкретной предметной области используются онтологические спецификации, задающие определения понятий предметной области и связей между ними. Онтологический контекст есть набор онтологических понятий и их связей, обеспечивающий правильную интерпретацию спецификаций в предметной области. Онтологические понятия *аннотируют* элементы спецификаций в посреднике и в информационных источниках.

Использование близких по смыслу онтологических спецификаций в элементах разных спецификаций является необходимой предпосылкой корректной взаимной интерпретации таких элементов спецификаций. Тем самым, онтологии представляют основу для семантического взаимодействия элементов спецификаций

Вербальное представление онтологий заключается в их определении на естественном языке (как в толковом словаре). При этом для установления позитивных связей и связей обобщения – специализации используется векторная модель информационного поиска. Поскольку онтологические понятия представляют собой сущности представления знаний, их структурные и логические свойства более формально выражаются в терминах абстрактных типов данных канонической модели. В этом случае связи между понятиями устанавливаются на основе отношения уточнения.

Онтологические спецификации используются для поиска классов и типов информационных источников, релевантных классам и типам посредника. Элемент спецификации источника *онтологически релевантен* элементу спецификации посредника того же вида (класс, тип, атрибут, функция, параметр), если между соответствующими им онтологическими понятиями установлена позитивная ассоциация, или ассоциация обобщения/специализации.

Поскольку спецификация посредника и каждый информационный источник могли разрабатываться с использованием разных онтологических спецификаций, возникает задача интеграции онтологий. Для этого необходимо установить ассоциации между онтологическими понятиями спецификации посредника и источника. Такие ассоциации устанавливаются при помощи набора специальных алгоритмов. Разработанный подход основан на интеграции онтологических контекстов компонентов и спецификации посредника, используя понятие общей онтологии предметной области. Определены алгоритмы, обеспечивающие указанную интеграцию, и поиск в таком интегрированном контексте спецификаций компонентов и их фрагментов, онтологически релевантных спецификации требований.

3.4. Устранение конфликтов

При регистрации в посреднике релевантных информационных источников неизбежно возникают различные конфликты между спецификациями посредника и источников. Конфликты могут возникать как из-за разных областей применения, так и из-за разного видения разработчиками представления спецификаций подобных друг другу объектов и классов. Применяемый способ разрешения конфликтов между спецификациями основан на комбинации двух подходов в области интеграции схем баз данных – применения набора предопределенных правил структурных преобразований, и применения языка высокого уровня для описания функций разрешения конфликтов. Функции разрешения

конфликтов задаются с помощью формул канонической модели. Язык формул является вариантом типизированного языка логики первого порядка. Для *разрешения конфликтов структурного вида* используются правила структурных преобразований. Они устанавливают релевантность путей в спецификациях типов посредника и типов информационных источников и задают правила построения функций разрешения конфликтов. Разработаны алгоритмы и программный инструментарий, позволяющие автоматизировать процесс поиска и разрешения структурных конфликтов между спецификациями посредника и источников.

3.5. Идентификация типов источников как уточнений типов посредника

Процесс реализации типов и классов посредника основан на выявлении фрагментов спецификаций существующих информационных источников и их дальнейшей композиции, уточняющей спецификацию посредника. Для этого используются операции над типами, ведущие к трансформации их спецификаций – операции декомпозиции и композиции. Процесс конструирования основан на понятии уточнения. *Уточняющие спецификации*, образуемые при конструировании, согласно теории уточнения, могут использоваться всюду *вместо уточняемых* спецификаций требований, так что пользователи не замечают этой замены. Методы уточнения позволяют формально устанавливать факт уточнения, гарантируя адекватность полученных конкретизирующих спецификаций требуемым.

Понятие *наибольшего общего редукта* является фундаментальным для этапа конструирования: оно составляет базис для определения повторно используемых фрагментов. Алгоритм конструирования наибольшего общего редукта состоит в следующем. Для определения наибольших общих редуктов для каждой пары онтологически релевантных типов *Ts* и *Tr* требуется найти максимальный набор *A* пар атрибутов (a_{Ts}, a_{Tr}), которые являются онтологически релевантными и имеют типы такие, что атрибут a_{Tr} можно использовать вместо a_{Ts} (для этого тип a_{Tr} должен уточнять тип a_{Ts}).

После выполнения этого этапа посредник готов к работе. Следует заметить, что регистрация новых источников может продолжаться и в дальнейшем при появлении новых релевантных посреднику компонентов.

4. Заключение

Таким образом, основными принципами организации решения задач над множественными источниками информации при ориентации на проблему являются следующие:

- независимость определения системы решения задачи (посредника) от существующих источников информации;

- определение посредника как результата консолидации усилий соответствующего научного сообщества;

- фокусирование на семантике и абстрактных определениях при спецификации посредника, что позволяет привлечь теоретиков к этапу его консолидации;

- независимость интерфейсов пользователей от используемых множественных информационных источников: пользователи посредника должны знать только определения предметной области в посреднике (определения понятий, структуры и

поведения объектов предметной области), благодаря чему они могут формулировать запросы при решении задач независимо от фактического набора информационных источников, зарегистрированных в посреднике;

- публикация информации о вновь разработанных источниках информации осуществляется в любое время и независимо от действующих к этому времени предметных посредников (в результате публикации могут быть инициированы действия по регистрации новых источников в посредниках, которым они релевантны);

- трехступенчатая идентификация релевантных посреднику информационных источников, обеспечивающая релевантность источников нефункциональным требованиям посредника (например, требованиям к качеству данных), онтологическую релевантность (установление соответствия онтологического контекста источника контексту посредника), структурная и поведенческая релевантность (доказательство структурного, функционального и процессного уточнения фрагмента спецификаций посредника фрагментом спецификаций источника);

- семантическая интеграция релевантных неоднородных информационных источников в посреднике;

- интегрированный доступ к информационным источникам, зарегистрированным в посреднике, при решении задач;

- рекурсивная структура посредников: каждый посредник регистрируется как новый информационный источник, что в частности, является полезным при решении задач на стыке различных предметных областей.

Инфраструктура системы организации решения задач над множественными информационными источниками принадлежит к классу информационных грид-архитектур. Она обеспечивает разработчиков стандартными интерфейсами для включения (plug-in) новых программных инструментов и баз данных одновременно с публикацией их метаданных стандартным образом. Инфраструктура реализуется в среде Web на основе Web сервисов. Основным исполнительным механизмом инфраструктуры является система управления потоками работ, позволяющая использовать в качестве их отдельных шагов вызовы произвольных сервисов наряду с запросами к базам данных и к посреднику. Инфраструктура содержит также средства поддержки реестров метаданных на основе протоколов Open Archive Initiative (OAI). Эта инфраструктура представляет собой оболочку, в которую встраиваются средства поддержки посредников периода исполнения. Эти средства включают репозиторий хранения метаинформации посредника, язык формулирования запросов к посреднику, средства переписывания запросов к посреднику в обращения к зарегистрированным в посреднике информационным источникам (используя их адаптеры), средства планирования реализации запросов в распределенной среде, средства управления посредниками. Сам посредник реализуется при этом как Web сервис.

Поскольку рассмотренный подход к решению задач ориентирован в первую очередь на применение в Российской Виртуальной Обсерватории [4], в качестве такой инфраструктурной оболочки планируется использование системы AstroGrid [19], разработанной в Великобритании и реализующей перечисленные функции.

5. Литература

1. J. -R. Abrial. The B-Book. Cambridge University Press, 1996

2. Briukhov D.O., Kalinichenko L.A. Component-based information systems development tool supporting the SYNTHESIS design method. Proceedings of the East European Conference on “Advances in Databases and Information Systems" (ADBIS'98), September 1998, Poland, Springer, LNCS N 1475, 1998

3. Briukhov D.O., Kalinichenko L.A., Skvortsov N.A. Information sources registration at a subject mediator as compositional development. Advances in Databases and Information Systems (ADBIS'01), Springer, Lecture Notes in Computer Science, 2151, 2001, pp 70 – 83

4. Briukhov D.O., Kalinichenko L.A., Zakharov V.N., Panchuk V.E., Vitkovsky V.V., Zhelenkova O.P., Dluzhnevskaya O.B., Malkov O.Yu., Kovaleva D.A Information Infrastructure of the Russian Virtual Observatory (RVO). Second Edition. IPI RAN, 2005

5. A.Y. Halevy. Answering queries using views: a survey. VLDB Journal, 10(4): 270 – 294, 2001

6. Калиниченко Л.А. Методы и средства интеграции неоднородных баз данных. – Москва: Наука, 1983. – 423 с.

7. Kalinichenko L.A. Methods and tools for equivalent data model mapping construction. EDBT’90 Conference: Proceedings. – Springer, 1990

8. Калиниченко Л.А. СИНТЕЗ: язык определения, проектирования и программирования интероперабельных сред неоднородных информационных ресурсов. – Москва: ИПИ РАН, 1993

9. Kalinichenko L.A. Method for Data Models Integration in the Common Paradigm Advances in Databases and Information Systems: Proceedings of the First East-European Conference. St. Petersburg, 1997

10. Kalinichenko L.A. Compositional Specification Calculus for Information Systems Development Proceedings of the East-West Conference on Advances in Databases and Information Systems (ADBIS'99), Maribor, Slovenia, September 1999, Springer Verlag, LNCS

11. L.A. Kalinichenko, N.A. Skvortsov Extensible ontological modeling framework for subject mediation In Proceedings of the 4-th Russian Scientific Conference "DIGITAL LIBRARIES: Advanced Methods and Technologies, Digital Collections, Oct. 15-17, 2002, Dubna

12. Kalinichenko L.A., Martynov D.O., Stupnikov S.A. Query rewriting using views in a typed mediator environment In Proceedings of the East-European Conference on “Advances in Databases and Information Systems" (ADBIS'04), Hungary, Budapest, Springer, Lecture notes in Computer Science, Vol. 3255, September 2004

13. Л.А. Калиниченко, С.А. Ступников, Н.А. Земцов. Синтез канонических моделей для интеграции неоднородных источников информации. Москва: ИПИ РАН, 2005

14. Uwe Keller, Rub´en Lara, Holger Lausen, Axel Polleres, and Dieter Fensel. Automatic Location of Services. Proceedings of ESWC 2005, Springer Verlag, LNCS 3532, pp. 1–16, 2005

15. Kosky A. Transforming Databases with Recursive Data Structures. PhD thesis, Department of Computer and Information Science, University of Pennsylvania, November 1995

16. Mili R., Mili A., Mittermeir R. Storing and retrieving software components: a refinement based systems. IEEE Transactions on Software Engineering, v. 23, N 7, July 1997

17. Skvortsov N. A., Kalinichenko L.A. An Approach to Ontological Modeling and Establishing Intercontext Correlation in the Semistructured Environment 2-nd Russian Scientific Conference "DIGITAL LIBRARIES: Advanced Methods and Technologies, Digital Collections, Sep. 26-28, 2000, Protvino

18. Stupnikov S.A. Mapping of Specification Canonical Model to Formal Notation for Refining Specifications Modelling. In Proceedings of the XXIV Conference of Young Scientists, Faculty of Mechanics and Mathematics, Moscow State University, April 8-13, 2002, Moscow

19. Walton, N. A., Lawrence, A., Linde, T. AstroGrid: Initial Deployment of the UK's Virtual Observatory, in ASP Conf. Ser., Vol. 314 Astronomical Data Analysis Software and Systems XIII, 2003

20. J. Wang, M.Maher, R. Topor. Rewriting Unions of General Conjunctive Queries Using Views. In Proc. of the 8th International Conference on Extending Database Technology, EDBT'02, Prague, Czech Republic, March 2002

21. Zaremski A.M., Wing J.M. Specification matching of software components. ACM Transactions on Software Engineering and Methodology, v. 6, N 4, October 1997