На правах рукописи

Привезенцев Алексей Иванович

ОРГАНИЗАЦИЯ ОНТОЛОГИЧЕСКИХ БАЗ ЗНАНИЙ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ ОПИСАНИЯ ИНФОРМАЦИОННЫХ РЕСУРСОВ В МОЛЕКУЛЯРНОЙ СПЕКТРОСКОПИИ

Специальность 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Томск – 2009

Работа выполнена в Институте оптики атмосферы СО РАН

Научный руководитель: кандидат физико-математических наук, старший научный сотрудник Фазлиев Александр Зарипович

Официальные оппоненты: доктор технических наук, профессор Янковская Анна Ефимовна;

кандидат технических наук, старший научный сотрудник Загорулько Юрий Алексеевич

Ведущая организация: Новосибирский государственный университет

Защита состоится «17» декабря 2009 г. в 10 час. 30 мин. на заседании диссертационного совета Д 212.267.08 по адресу: 634050, г. Томск, пр. Ленина, 36, корп. 2, ауд. 102, Томский государственный университет.

С диссертацией можно ознакомиться в научной библиотеке Томского государственного университета по адресу: 634050, г. Томск, пр. Ленина, 34а.

Отзывы на автореферат (2 экз.), заверенные печатью, высылать по адресу: 634050, г. Томск, пр. Ленина, 36, ученому секретарю ТГУ.

Автореферат разослан 16 ноября 2009 г.

Ученый секретарь
диссертационного совета,
д.т.н., профессор А.В. Скворцов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы диссертационной работы.

Молекулярная спектроскопия является одним из широко используемых во многих прикладных исследованиях разделов физики. Предметом изучения молекулярной спектроскопии являются спектральные свойства молекул. Детальное изучение спектральных свойств молекул не закончено до сих пор. Связано это с тем обстоятельством, что в расчетах физических характеристик атмосферы используются сотни тысяч линий, каждая из которых описывается десятком параметров. В молекулярной спектроскопии постоянно публикуется огромное количество сложных результатов измерения или расчетов спектров – результаты решения предметных задач. Решаются предметные задачи для расчёта сотен миллионов линий, проводятся эксперименты с помощью современной техники для измерения спектров, которая позволяет получать данные с большей точностью и в тех диапазонах длин волн, в которых ранее измерения не проводились. Также растет число исследовательских групп. Кроме увеличения объёма спектральных данных постоянно меняется структура представления данных, как с предметной точки зрения так с технической реализации. Например, за почти сорокалетнюю историю одна из ведущих групп экспертов по спектроскопии, поддерживающая базу данных HITRAN, несколько раз модифицировала как набор физических сущностей, так и формат документов и файлов, в котором хранятся данные. Все это указывает на необходимость сбора, хранения, обработки и распространения информации с использованием современных подходов для коллективной работы на базе информационных систем в сети Internet.

На данный момент, для работы с этой информацией создаются специальные базы постоянно пополняющихся спектральных данных: HITRAN, GEISHA, VALD, CDMS, BASECOL, STSP. Работа с такими массивами данных требует, с одной стороны, предметной систематизации данных, с другой стороны, программных средств для их автоматизированной обработки, включающей программную интеграцию и структурирование разнородных ресурсов из различных предметных областей, а также возможность подготовки данных для решения прикладных задач в смежных предметных областях: астрономии, атмосферной радиации, оптики атмосферы. Поэтому, на основе этих баз данных создаются информационные системы. Несмотря на это, для молекулярной спектроскопии характерны следующие информационные проблемы, не решаемые в существующих информационных системах:

- для коллективной работы в информационной системе у пользователя отсутствует возможность самостоятельного формирования структуры массивов спектральных данных и их наполнение конкретными значениями, проведения на их основе расчетов и сравнения с результатами экспериментов;

- базы спектральных данных могут содержать недостоверные данные, что снижает их научную ценность;

- имеется неопределенность в информации о собранных данных, об их способах получения;

- существующие информационные системы не дают средств для автоматизированного программного анализа информации о данных и её последующей логической машинной обработки, необходимой для построения Semantic Web.

Идея Semantic Web состоит в машинной логической обработке семантики информационных ресурсов, имеющихся в сети Internet, для автономного решения интеллектуальных задач. Для решения таких задач должны использоваться специализированные интеллектуальные программы-агенты, которые предлагают решения, используя базу знаний, основанную на онтологии (*онтологическую базу знаний*). Для организации онтологий консорциум W3C, разрабатывающий Semantic Web, определил в качестве спецификации язык OWL DL.

Активные исследования по представлению знаний в виде онтологии начались в начале 1990-х и продолжаются до сих пор. Среди большого количества работ можно выделить M.R. Genesereth, T.R. Gruber, N. Guarino, R. Mizogushi, J.F. Sowa, R. Studer. Актуальные исследования онтологий в рамках Semantic Web представлены в работах I.A. Horrocks, D.L. McGuinness, P.F. Patel-Schneider. Среди отечественных публикаций существует разнообразие подходов к представлению знаний в виде онтологий, и исследования в данной области активно ведутся И.Л. Артемьевой, Е.М. Бенеаминовым, В.И. Воробьевым, Б.В. Добровым, Т.А. Гавриловой, Н.Г. Загоруйко, Ю.А. Загорулько, Л.А. Калиниченко, А.С. Клещевым, Н.В. Лукашевич, Д.Е. Пальчуновым, А.Ф. Тузовским, В.Ф. Хорошевским. Большое количество публикаций в данной области указывает на решение разнообразных задач с помощью баз знаний, основанных на онтологиях.

Онтологические базы знаний позволяют осуществлять открытое представление машинно-обрабатываемых знаний, что позволяет повысить эффективность коллективной работы ученых в своих узкоспециализированных предметных областях. Так как они дают возможность учёным строить собственные концептуализации предметной области и проверять согласованность своих знаний с другими экспертными публикуемыми знаниями. Кроме этого ученые, анализируя получаемое

знание о результатах решения предметных задач, могут своевременно реагировать на важные сведения, например о некорректных данных. Кроме того, использование онтологических баз знаний для описания разнородных данных в рамках научных информационно-вычислительных систем позволяет решать задачи классификации, интеграции, поиска и сравнения информационных ресурсов.

В настоящее время в молекулярной спектроскопии в рамках научных информационно-вычислительных систем отсутствуют машинно-обрабатываемые базы знаний.

На основе всего вышеперечисленного можно сделать вывод о том, что исследование подхода к организации онтологической базы знаний по молекулярной спектроскопии имеет научную и практическую актуальность.

Цель диссертационной работы: разработка и исследование подхода к построению в рамках научной информационно-вычислительной системы онтологических баз знаний для описания разнородных данных молекулярной спектроскопии, извлечённых из научных публикаций и проверяемых на достоверность.

Для достижения цели диссертационной работы решаются следующие **задачи**:

1. Создание информационных моделей для представления данных и знаний в области молекулярной спектроскопии.

2. Разработка структур данных для информации, извлеченной из научных публикаций по спектроскопии молекул, допускающих автоматическую проверку целостности данных и необходимых для обмена между интеллектуальными агентами.

3. Создание терминологической компоненты (TBox) онтологической базы знаний для представления знаний в области молекулярной спектроскопии.

4. Разработка алгоритма формирования онтологического описания опубликованных данных с целью построения набора фактов в базе знаний по молекулярной спектроскопии, содержащих знания о их первоисточниках и достоверности.

5. Реализация программного обеспечения, созданного на основе разработанного алгоритма онтологического описания информационных ресурсов и практического использования этого описания в НИВС по спектроскопии молекул воды.

6. Реализация фактографической компоненты (ABox) онтологической базы знаний по описанию опубликованных данных спектроскопии молекул воды.

Объектом исследования являются структуры данных и модели представления знаний в информационных системах по молекулярной спектроскопии.

Предметом исследования являются подходы и алгоритмы создания баз знаний и систем управления ими в научных информационно-вычислительных системах по молекулярной спектроскопии.

Методы исследования. В ходе диссертационного исследования были использованы методы онтологического моделирования, теории множеств, дескриптивной логики, объектно-ориентированного проектирования и программирования.

Научная новизна диссертационной работы заключается в следующем:

1. Впервые построена семантическая модель в виде терминологической компоненты (TBox) базы знаний, являющаяся объединением информационных моделей объектов молекулярной спектроскопии, представляющая собой решения двух цепей прямых и обратных задач спектроскопии и свойств решений этих задач, позволившая решить задачу автоматической систематизации знаний о достоверности этих решений.

2. Впервые создан алгоритм для автоматизации построения фактологической компоненты (ABox) базы знаний о решениях задач молекулярной спектроскопии и их свойствах, являющийся необходимым для машинной актуализации знаний о достоверности решений задач и применимый для всех спектральных молекул.

3. Впервые создана онтологическая база знаний спектроскопии молекул воды, в которой фактологическая компонента (ABox) содержит наиболее полную информацию о значениях параметров спектральных линий молекул воды, опубликованную в мире.

Теоретическая значимость исследования состоит в разработке онтологии спектроскопии молекул как основы для построения и проверки научных гипотез, разнообразных систематизаций знаний, интеграции знаний различных предметных областей, что открывает перспективы для постановки и решения новых предметных задач, как в молекулярной спектроскопии, так и смежных с ней областях науки, таких как астрономия, атмосферная радиация, оптика атмосферы.

Практическая ценность диссертационной работы заключается:

1. В создании наиболее полной прикладной онтологии по опубликованным данным спектроскопии молекул воды.

2. В возможности использования открытых результатов семантического описания решений задач, оформленных по стандарту OWL DL, во внешних специализированных системах по работе с онтологиями, использующих машины вывода.

3. В разработке программного обеспечения в рамках научной информационно-вычислительной системы, имеющей трёхслойную архитектуру, на основе предложенного алгоритма онтологического описания информационных ресурсов и применении этого программного обеспечения рядом ведущих исследовательских групп спектроскопистов в России (Санкт-Петербургский государственный университет, Институт прикладной физики РАН, Институт оптики атмосферы РАН).

Основные защищаемые положения:

1. Семантическая модель в виде терминологической компоненты (TBox) базы знаний, являющаяся объединением информационных моделей объектов молекулярной спектроскопии, представляющая собой решения двух цепей прямых и обратных задач спектроскопии и свойств решений этих задач.

2. Алгоритм для автоматизации построения фактологической компоненты (ABox) базы знаний о решениях задач молекулярной спектроскопии и их свойствах.

3. Программное обеспечение в рамках научной информационно-вычислительной системы, имеющей трёхслойную архитектуру, созданное на основе разработанного алгоритма онтологического описания информационных ресурсов и полученная с его использованием онтологическая база знаний спектроскопии молекул воды.

Апробация диссертационной работы. Все результаты работы докладывались и обсуждались на следующих научных конференциях: IX Рабочем совещании по электронным публикациям «El-Pub2004» – Новосибирск, 23-25 сентября 2004; V Всероссийской конференции молодых ученых по математическому моделированию и информационным технологиям – Новосибирск, 1-3 ноября 2004; Международной конференции по вычислительно-информационным технологиям для наук об окружающей среде «Cites-2005» – Новосибирск, 13-23 марта 2005; X Байкальской Всероссийской конференции «Информационные и математические технологии в науке, технике и образовании» – Северобайкальск, 12-19 июля 2005; 7-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL’2005) – Ярославль, 4-6 октября 2005; International conference on environment observations, modeling and informational systems (ENVIROMIS-2006) – Tomsk, 1-8 June 2006; XVth Symposium on High Resolution Molecular Spectroscopy «HighRus-2006» – Nizhny Novgorod, 18-21 July 2006; Рабочем семинаре «Проблемы и решения задач в области наук о Земле в распределенной ИНТЕРНЕТ среде» – Москва, 13-15 февраля 2007; European Geosciences Union General Assembly 2007 – Vienna, 15-20 April 2007; International conference

on Computational Information Technologies for Environmental Sciences «Cites-2007» – Томск, 14-25 июля 2007; Всероссийской конференции «Знания – Онтологии – Теория» – Новосибирск, 14-16 сентября 2007; 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2007) – Переславль-Залесский, 15-18 октября 2007; Всероссийской научно-практической конференции «Свободное программное обеспечение: разработка и внедрение» - Томск, 17-18 мая 2008; XIII Байкальской всероссийской конференции «Информационные и математические технологии в науке, технике и образовании» – Иркутск, 7-16 июля 2008; International conference on environment observations, modeling and informational systems (ENVIROMIS-2008) – Tomsk, 28-5 July 2008; European geosciences union general assembly 2009 – Vienna, 19-25 April 2009; XVI Международном симпозиуме «Оптика атмосферы и океана. Физика атмосферы» – Томск, 12-15 октября 2009; IV Всероссийской конференции молодых учёных «Материаловедение, технологии и экология в 3-м тысячелетии» – Томск, 19-21 октября 2009; Всероссийской конференции «Знания – Онтологии – Теория» – Новосибирск, 20-22 октября 2009.

По теме диссертационной работы ***опубликовано 17 научных работ***:

- из них шестнадцать печатных [1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17] и одна в электронном журнале [5];
- из них четырнадцать работ на русском языке [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17] и три на английском [1, 2, 3];
- из них две в журналах из перечня ВАК по управлению, вычислительной технике и информатике [4, 9], две в журналах из перечня ВАК по физике [8, 11], две в журналах [5, 10], одиннадцать в трудах и материалах конференций [1, 2, 3, 6, 7, 12, 13, 14, 15, 16, 17].

Внедрение результатов диссертационной работы, было осуществлено в трех основных исследовательских группах спектроскопистов в России:

- Институт оптики атмосферы СО РАН, где результаты доступны активно используется в рамках НИВС (http://saga.iao.ru);
- Институт прикладной физики РАН, где результаты доступны в рамках НИВС по адресу http://saga.atmos.appl.sci-nnov.ru;
- Санкт-Петербургский государственный университет, где результаты доступны в рамках НИВС по адресу http://saga.molsp.phys.spbu.ru.

Работа выполнена ***при поддержке грантов***: Российского Фонда Фундаментальных Исследований (РФФИ) «Распределенная информационная система «Молекулярная спектроскопия»« (05-07-90196, А.Д.

Быков, 2005-2007); РФФИ «Интернет доступная информационная система по молекулярной спектроскопии, основанная на знаниях» (08-07-00318-а, А.З. Фазлиев, 2008-2010); UIPAC task 2004-035-1-100 «A database of water transitions from experiment and theory».

Личный вклад автора.

Опубликованные работы написаны в соавторстве с экспертами предметной области спектроскопии молекулы воды и сотрудниками центра интегрированных информационных систем ИОА СО РАН. В совместных работах диссертант принимал участие в непосредственной разработке схем XML-данных, метаданных и их дальнейшем внедрении в модель НИВС, в разработке прикладной онтологии задач по спектроскопии молекул воды, во внедрении результатов работы. В разработке перечисленного программного обеспечения ему принадлежит определяющая роль.

Благодарности.

Автор выражает благодарность профессорам А.А. Мицелю и А.Ф. Тузовскому за внимание к работе, ценные замечания и помощь, способствующие окончательному варианту рукописи. Автор признателен с.н.с. А.З. Фазлиеву за ценные консультации, постановки задач и всестороннюю поддержку данной работы.

Автор благодарен чл.-корр. РАН С.Д. Творогову, а также благодарит д.ф.-м. н. А.Д. Быкова и к.ф.-м.н. Б.А. Воронина за консультации и помощь при определении структуры данных в молекулярной спектроскопии воды; Н.А. Лаврентьева за реализацию программ для расчета коэффициентов поглощения газов; А.Ю. Ахлёстина за реализацию ядра НИВС; А.В. Козодоева за реализацию системы ввода данных; д.ф.-м.н. О.Б. Родимову за помощь в составлении типовых вопросов для задачи нахождения уровней энергии молекулы.

Структура и объём диссертационной работы. Диссертация состоит из перечня условных обозначений, введения, четырёх глав, заключения, списка использованных источников и шести приложений. Общий объём работы составляет 239 страниц. Список использованных источников насчитывает 128 наименований. Работа содержит 42 рисунка и 25 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении излагается положение дел в прикладной части молекулярной спектроскопии, обосновывается актуальность темы диссертационной работы, сформулированы цель и задачи работы, изложены основные научные результаты, выносимые на защиту, и их практическая значимость.

В **первой главе** излагается подход к организации онтологических баз знаний в информационных системах.

Определение. *Знания* – совокупность зафиксированных в сознании и мышлении человека или машины фактов предметной области. Сознание определяет восприятие и понимание окружающего мира, а мышление задаёт способы установления связей, сопоставлений и осуществление из этого выводов, используя логику. Если рассматривать логику, заложенную в машину, то используется формальный язык математической логики. Формальный язык (множество конечных слов над конечным алфавитом) определяется словами, порождёнными некоторой формальной грамматикой, например формой Бэкуса-Наура для описания синтаксиса. Интерпретации смысла синтаксических конструкций формального языка логики определяется логической семантикой, которая определяется формально. Формализация логической семантики позволяет явно задавать смысл высказываний и разделять сами высказывания и заключения об их истинности.

Для достижения компромисса между логикой предикатов, имеющей формальную семантику, и семантическими сетями, являющимися удобным способом представления знаний предметной области в виде иерархий понятий и их отношений, используется семейство дескриптивных логик, которое является подмножеством логики предикатов первого порядка.

Во вводной части главы рассматриваются разные точки зрения на понятие «онтология». Рассматриваются подходы к онтологической концептуализации и различные классификации онтологий. Представлено описание инженерии онтологий, языков онтологий и инструментальных средств для работы с ними.

В диссертационной работе понятие «онтология» используется в соответствии с определением языка спецификации онтологий OWL, имеющим формальный синтаксис и формальную семантику семейства дескриптивных логик, например, OWL Lite имеет основу в виде дескриптивной логики под названием SHIF(D), а OWL DL – SHOIN(D).

Определение. *Онтология* (*база знаний*) – совокупность TBox и ABox на формальном языке OWL DL.

Формальный синтаксис языка содержит алфавит, состоящий из трёх компонентов $\{C,\ R,\ I\}$, где $C = \{T,\ \perp,\ C_1...C_m\}$ – конечное множество имён классов (понятий), включающее универсальный класс T и пустой класс $\perp$; $R = \{R_1...R_k\}$ – конечное множество имён свойств (бинарных отношений); $I = \{i_1...i_n\}$ – конечное множество имён объектов (экземпляров). Набор *терминологических аксиом* типа «дочерний класс», «эквивалентный класс», «дочерний свойство», «эквивалентное свойство» называется TBox (сокращение от terminological box), то есть вводит терминологию предметной области. Набор *аксиом утверждений* типа «экземпляр класса», «экземпляр свойства» называется ABox (assertional box), то есть содержит утверждения об именованных индивидах в заданной терминологии.

Определение. *Семантические метаданные* – описание информационного ресурса (ABox) относительно некоторой терминологии онтологической модели предметной области (TBox).

В работе семантические аннотации представлены с помощью языка спецификации OWL DL. Такие аннотации будем называть *онтологическими аннотациями* или *онтологическими описаниями*. Процесс формирования структуры семантической аннотации обусловлен задачей перевода интерпретируемого человеком описания в описание, интерпретируемое машиной.

В главе представлен обзор существующих НИВС по молекулярной спектроскопии.

Во второй главе рассмотрен подход к информационной модели предметной области в виде цепей её прямых и обратных задач.

В естественных науках большая часть задач связана со *знаниями*, *основанными на решениях задач предметной области*. Целью решения предметной задачи является изучение состояний физической системы. Эти состояния исследуемой системы при представлении знаний рассматриваются как наборы фактов (ABox). Задачи классификации в молекулярной спектроскопии, как правило, сводятся к построению таксономии терминов предметной области и на практике рассматриваются как вспомогательные задачи. Предполагается, что в задаче классификации концепты представляют интенсионалы предметной области (TBox). Описание предметной области основано на нескольких таксономиях и наборах фактов для каждой из задач модели предметной области.

Важную роль при создании базы знаний играет концепт *«информационный источник»*. Концепт *информационный источник* описывает решение задачи предметной области и важен при семантическом описании решений задач молекулярной спектроскопии.

Основной акцент для модели предметной области в виде цепей прямых и обратных задач сделан на *автоматическое установление достоверности данных решений задач*. В проверке на достоверность результатов решений задач можно выделить несколько уровней:

1. Проверка ограничений на типы данных.

2. Проверка ограничений на допустимые интервальные значения физических величин.

3. Проверка ограничений, следующих из математических моделей исследуемых физических объектов.

4. Систематизация результатов решений задач по величине среднеквадратичных отклонений.

5. Указание недостоверности данных, выявленных экспертом предметной области на основе «дальних корреляций» результатов решений задач.

Прямые задачи молекулярной спектроскопии связаны с расчетами из первых принципов фундаментальных характеристик молекул. Обратные задачи молекулярной спектроскопии связаны с обработкой данных измерений спектральных функций, что позволяет в дальнейшем при машинной обработке классифицировать их выходные данные как экспериментальные. В цепи задач молекулярной спектроскопии существуют связи между прямыми и обратными задачами.

При решении задач обоих типов проводятся вычисления одних и тех же физических величин. Их сравнение между собой позволяет делать выводы о достоверности данных.

К классам *элементарных прямых задач*, используемых нами для проектирования информационной системы, относятся следующие классы задач: задача определения физических характеристик изолированной молекулы (T1); задача определения параметров спектральной линии изолированной молекулы (T2); задача определения параметров контура спектральной линии (T3); задача расчета спектральных функций (T4).

К классам *элементарных обратных задач*, используемых нами для проектирования информационной системы, относятся следующие классы: задача измерения спектральных функций (E1); задача приписывания квантовых чисел спектральным линиям (T5); задача определения коэффициентов Эйнштейна (T6); задача определения уровней энергии изолированной молекулы (T7).

В работе для первых двух уровней автоматической проверки на достоверность данных решений задач можно использовать представление структурированных данных в виде XML-документа с соответствующей XML-схемой, с заданными ограничениями на типы данных и интервальные значения. В НИВС можно использовать проверку дан-

ных по XML-схемам для выявления ошибок загрузки и выделять данные, которые не соответствуют заданным ограничениям.

В рамках диссертационной работы созданы XML-схемы. Организация XML-схем основана на предположении, что любая физическая задача состоит в изучении определенной молекулы. Поэтому, корневым элементом в документе будет название изотопомера молекулы, ему соответствуют название файла содержащего XML-схему.

В третьей главе рассмотрен подход к построению базы знаний по спектроскопии молекулы воды, описан процесс автоматизации процедуры наполнения онтологии молекулярной спектроскопии новыми фактами. Созданные онтологии используются для машинной систематизации и интерпретации знаний, для интеграции знаний в другие смежные предметные области, а также организации семантического поиска. На рис 1. представлены базовые классы прикладной онтологии по молекулярной спектроскопии, представленные в нотации Protégé.

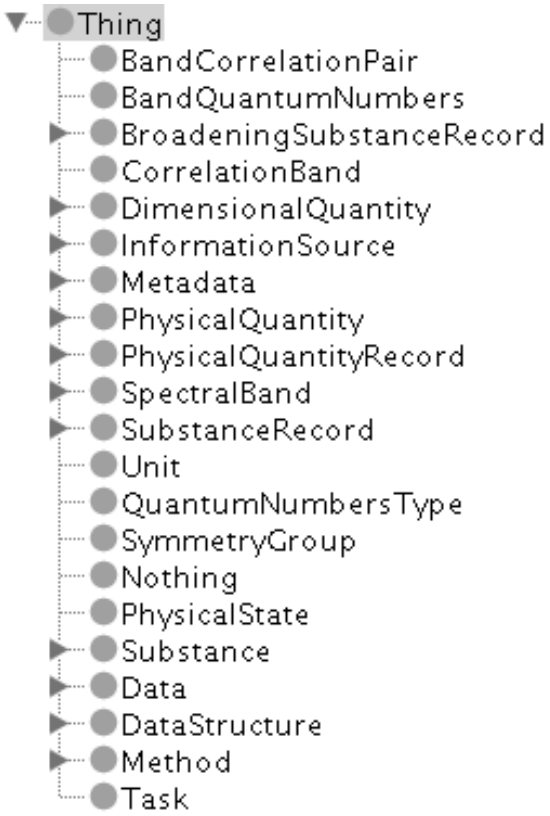

Рис. 1. Базовые классы онтологии по молекулярной спектроскопии

В этой онтологии можно выделить три группы классов:

Классы, содержащие объекты, относящиеся к спектроскопии молекул, в частности, молекулы воды и ее изотопомеров (**Substance**, **PhysicalState, PhysicalQuantity, DimensionalQuantity, QuantumNum-**

bersType, Unit, SpectralBand, BandQuantumNumbers, BandQuantumNumbers, CorrelationBand)

Классы, содержащие математические модели объектов, используемые в молекулярной спектроскопии (**SymmetryGroup, Task, Method**).

Классы, содержащие информационные объекты (**Metadata, InformationSource, SubstanceRecord, PhysicalQuantityRecord, BroadeningSubstanceRecord**).

Реализация онтологий по молекулярной спектроскопии проведена на языке OWL DL. Большая часть классов прикладной онтологии задач построены с помощью ограничений на свойства. Программная реализация TBox онтологии содержит 11 файлов, представляющих разные части предметной области.

Особенностью работы с онтологиями задач в НИВС «Молекулярная спектроскопия» является тот факт, что пользователи при решении задач механически составляют свою собственную онтологию, содержащую индивиды, соответствующие конкретной решенной задаче. Эти индивиды можно объединять с онтологиями других задач или других пользователей, если это позволяют права доступа к базе знаний.

Использование прикладной онтологии возможно на клиентском месте с помощью любого редактора онтологий, например Protégé. Средства этого редактора позволяют пользователю составлять запросы с помощью конструкции утверждения (субъект, предикат и объект), экземпляры которого можно создавать из концептов прикладных онтологий.

Полная прикладная онтология по молекулярной спектроскопии в НИВС «Молекулярная спектроскопия» содержит 1 347 796 утверждений на языке OWL. Из которых, 1 467 утверждений приходится на таксономию. На факты по задаче T1 приходится 1 432 утверждений. На факты по задаче T2 приходится 1 586 утверждений. На факты по задаче T3 приходится 4 265 утверждений. На факты по задаче T5 приходится 28 723 утверждений. На факты по задаче T6 приходится 17 202 утверждений. На факты по задаче T7 приходится 6 159 утверждений. На факты по корреляциям результатов решений задач T1 и T7, включая информацию в полосах, приходится 216 567. На факты по корреляциям результатов решений задач T2 и T6, включая информацию в полосах, приходится 543 515. На факты по корреляциям результатов решений задач T3 и T5, включая информацию в полосах, приходится 521 269.

В четвёртой главе описаны алгоритмы и программное обеспечение Мета+ для организации онтологической базы знаний.

Представлена общая структура программного обеспечения Мета+ в рамках существующей НИВС «Молекулярная спектроскопия», которая включает в себя:

- онтологическую базу знаний, хранящую метаданные и онтологии НИВС;
- прикладные программные интерфейсы POWL (модифицированный) и RAP, созданные сторонними разработчиками и предоставляющие функции доступа к онтологическому хранилищу;
- классы и библиотеки функций, составляющие ядро Мета+ и выполняющие задачи формирования, хранения, обработки и визуализации онтологического описания;
- конфигурационные файлы и шаблоны Мета+, позволяющие настраивать и организовывать работу различных функциональных блоков.

Кроме перечисленных блоков в программное обеспечение Мета+ входят модули и шаблоны, реализуемые по правилам, предлагаемым ядром НИВС:

- модули для взаимодействия с внешней средой через ядро НИВС, организующие web-ориентированное интерактивное взаимодействие с пользователем;
- шаблоны для формирования визуализации функциональности, реализуемой модулями.

В работе описываются алгоритмы формирования семантического описания на основе онтологии задач по молекулярной спектроскопии.

Алгоритм создания метаданных описывает поход к составлению метаданных для решений задач, используя анализ входных и выходных данных задач и способов их описания. Описание решений задач можно представить количественными и качественными метаданными.

Таким образом, имеем два способа получения метаданных. Первый способ – это автоматизированное генерирование метаданных непосредственно из доступных массивов данных. Будем называть такие метаданные вычисляемыми. Второй способ – это ввод метаданных пользователем, что требует создания дополнительных интерфейсов ввода. Будем называть такие метаданные невычисляемыми.

Шаг формирования вычисляемых метаданных отделен от шага заведения невычисляемых метаданных. На любом шаге можно создать экземпляр OWL-класса решения задачи, но этот индивид будет недостоверным в силу своей неполноты, в соответствии с требованиями из определения OWL-класса решения задачи. Индивид решения задачи

будет полным и достоверным, если будет выполнена последовательность из шагов по формированию вычисляемых и невычисляемых метаданных.

Алгоритм формирования индивидов онтологии предназначен для преобразования информации в семантические метаданные, основанные на прикладной онтологии решений задачи. Алгоритм состоит из систематизированной последовательности следующих шагов:

1. Создание онтологического уровня (TBox) прикладной онтологии решения задачи предметной области, описывающей входные и выходные данные решения задачи, а также дополнительные характеристики, полученные в результате применения алгоритма фиксации схем метаданных.

2. Создание шаблонного индивида OWL-класса решения задачи.

3. Создание на основе XML-синтаксиса шаблонного индивида OWL-класса решения задачи XML-шаблона для генератора индивидов, созданного по определенным правилам, накладываемым программной реализацией генератора.

4. Применение алгоритма создания метаданных, основанного на выявлении вычисляемых и невычисляемых метаданных, реализуя:

4.1. функции преобразования данных из конкретных решений задачи в метаданные (вычисляемые метаданные) для них;

4.2. интерфейсы запроса и функции обработки невычисляемых метаданных для решений задачи;

5. Формирование дополнительных программных модулей для использования в единой научной информационно-вычислительной системе.

Данная последовательность шагов является универсальной для составления индивидов OWL-классов решений задач.

В диссертационной работе представлено описание функциональности ядра программного обеспечения Мета+, предлагающего два направления разработки:

- функциональность для разработчика, включающая в себя существующие эффективные техники программирования, в частности, предварительное тестирование, использование исключений, интерфейсы и модули по поддержке разработки. В работе представлены некоторые UML-диаграммы тестов. Для сообщения об определенных видах ошибок были созданы новые классы исключений. Для разработчика предусмотрены отдельные интерфейсы для управления результатами своих действий, в частности модуль обновления для пересчета данных и занесения их в метаданные.

- функциональность для конечного пользователя, включающая в себя: функции визуализации массивов данных с учетом их метадан-

ных; функции экспортирования OWL/RDF моделей из НИВС во внешнюю среду в XML-синтаксисе; функции визуализации в HTML-формате индивидов OWL-классов решений задач; функции для составления онтологических экземпляров классов задач, используя HTML-формы и их обработку; функции для составления онтологических экземпляров классов задач, используя автоматический обсчёт имеющихся массивов данных; функции осуществления семантического поиска, используя интерфейс составления вопросов в виде связанных триплетов для получения адекватного ответа. В работе представлены некоторые UML-диаграммы классов ядра Мета+.

В диссертационной работе описана реализация программного обеспечения Мета+, осуществленного на языке PHP 5. Программное обеспечение Мета+ использует открытое свободное API для работы с RDF и OWL. Для работы с RDF используется RDF API for PHP (RAP), а для работы с OWL используется API pOWL, основанное на API RAP.

Основным языком для разметки информации в ПО Мета+ использовался язык XML. На этом языке создавались конфигурационные файлы, шаблоны для генерации онтологических экземпляров. Сами онтологии написаны в XML-синтаксисе RDF/XML.

В заключении сформулированы основные результаты работы и представлены основные направления развития представляемой работы.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Bykov, A. D. Distributed information system on atmospheric spectroscopy / A. D. Bykov, A. Z. Fazliev, N. N. Filippov, A. V. Kozodoev, **A. I. Privezentsev**, L. N. Sinitsa, M. V. Tonkov, M. Yu. Tretyakov // Geophysical Research Abstracts. European Geosciences Union General Assembly 2007 – Vienna, 15-20 april 2007. – Vol. 9. – Vienna: Copernicus, 2007. – 8 pp.

2. Bykov, A. D. Distributed information system on molecular spectroscopy / A. D. Bykov, A. Z. Fazliev, A. V. Kozodoev, **A. I. Privezentsev**, L. N. Sinitsa, M. V. Tonkov, N. N. Filippov, M. Yu. Tretyakov // Proc. of SPIE, 15th Symposium on High-Resolution Molecular Spectroscopy –Vol. 6580. – pp. 65800W. – 2006. – 12 pp.

3. Fazliev, A. Z. Semantic metadata application for information resources systematization in water spectroscopy / A. Z. Fazliev, **A. I. Privezentsev**, J. Tennyson // Geophysical Research Abstracts. European Geosciences Union General Assembly 2009 – Vienna, 19-25 april 2009. – Vol. 11. – Vienna: Copernicus, 2009. – 4 pp.

4. Быков, А. Д. Структурирование ресурсов информационной системы по молекулярной спектроскопии / А. Д. Быков, А. В. Козодоев, **А. И. Привезенцев**, А. З. Фазлиев // Вычислительные технологии. – 2007. – Т. 12. – С. 10-18.

5. Козодоев, А. В. Аннотирование информационных ресурсов в распределенной информационной системе «Молекулярная спектроскопия» [Электронный ресурс] / А. В. Козодоев, **А. И. Привезенцев**, А. З. Фазлиев // Электронные библиотеки: Российский научный электронный журнал. – Электрон. журн. – М.: Институт развития информационного общества, 2006. – Т. 9. – № 3. – Режим доступа: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2006/part3/KPF, свободный.

6. Козодоев, А. В. Аннотирование информационных ресурсов в распределенной информационной системе «Молекулярная спектроскопия» / А. В. Козодоев, **А. И. Привезенцев**, А. З. Фазлиев // Труды 7-ой Всероссийской конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2005) – Ярославль, 4-6 октября 2005. – Ярославль: Издательство Ярославского Государственного Университета, 2005. – С. 80-86.

7. Козодоев, А. В. Данные и метаданные в распределенной информационно-вычислительной системе «Молекулярная спектроскопия» / А. В. Козодоев, **А. И. Привезенцев**, А. З. Фазлиев // Труды X Байкальской Всероссийской конференции «Информационные и математические технологии в науке, технике и образовании» – Северобайкальск, 12-19 июля 2005. – Ч. 1. – Иркутск: Издательство ИСЭМ СО РАН, 2005. – С. 45-50.

8. Козодоев, А. В. Информационная система для решения задач молекулярной спектроскопии. 3. Уровни энергии молекул / А. В. Козодоев, **А. И. Привезенцев**, А. З. Фазлиев // Оптика атмосферы и океана. – 2007. – Т. 20. – № 9. – С. 805-809.

9. Козодоев, А. В. Организация информационных ресурсов в распределенной информационно-вычислительной системе, ориентированной на решение задач молекулярной спектроскопии / А. В. Козодоев, **А. И. Привезенцев**, А. З. Фазлиев // Вычислительные технологии. – 2005. – Т.10., спец. выпуск.– С. 82-91.

10. Козодоев, А. В. Структура ресурсов информационно-вычислительной системы по молекулярной спектроскопии / А. В. Козодоев, **А. И. Привезенцев**, А. З. Фазлиев // Измерения, моделирование и информационные системы для изучения окружающей среды : сб. ст. / под ред. Е. П. Гордова. – Томск: Томский ЦНТИ, 2006. – С. 32-35.

11. Лаврентьев, Н. А. Информационная система для решения задач молекулярной спектроскопии. 4. Переходы в молекулах симмет-

рии C2v и Cs / Н. А. Лаврентьев, **А. И. Привезенцев**, А. З. Фазлиев // Оптика атмосферы и океана. – 2008. – Т. 21. – № 11. – С. 957-962.

12. Лаврентьев, Н. А. Распределенная информационная система по молекулярной спектроскопии углекислого газа / Н. А. Лаврентьев, **А. И. Привезенцев**, А. З. Фазлиев // Материалы XVI Международного симпозиума «Оптика атмосферы и океана. Физика атмосферы» – Томск, 12-15 октября 2009. – Томск: Издательство ИОА СО РАН, 2009. – С. 42-45.

13. Привезенцев, А. И. Прикладная онтология для задач молекулярной спектроскопии / **А. И. Привезенцев**, А. З. Фазлиев // Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2007) – Переславль-Залесский, 15-18 октября 2007. – Т. 1. – Переславль-Залесский: Издательство института программных систем РАН, 2007. – 10 с.

14. Привезенцев, А. И. Прикладная онтология задач для молекулярной спектроскопии / **А. И. Привезенцев**, А. З. Фазлиев // Труды Всероссийской конференции «Знания – Онтологии – Теория» – Новосибирск, 14-16 сентября 2007. – Т. 2. – Новосибирск: Омега Принт, 2007. – С. 82-87.

15. Привезенцев, А. И. Применение семантических метаданных для систематизации информационных ресурсов в молекулярной спектроскопии / **А. И. Привезенцев**, А. З. Фазлиев // Труды XIII Байкальской Всероссийской конференции «Информационные и математические технологии в науке, технике и образовании» – Иркутск, 7-16 июля 2008. – Ч. 1. – Иркутск: Издательство ИСЭМ СО РАН, 2008. – С. 171-176.

16. Привезенцев, А. И. Логическая молекулярная спектроскопия / **А. И. Привезенцев**, А. З. Фазлиев, J. Tennyson // Материалы Всероссийской конференции «Знания – Онтологии – Теории» – Новосибирск, 20-22 октября 2009. – Т. 2. – Новосибирск: РИЦ Прайс-Курьер, 2009. – С. 202-206.

17. Привезенцев, А. И. Онтологическая база знаний по описанию результатов решений задач в молекулярной спектроскопии / **А. И. Привезенцев** // Материалы IV Всероссийской конференции молодых учёных «Материаловедение, технологии и экология в 3-м тысячелетии» – Томск, 19-21 октября 2009. – Томск: Издательство ИОА СО РАН, 2009. – С. 624-628.

Тираж отпечатан в типографии ИОА им. В.Е. Зуева СО РАН
634021, г. Томск, пл. Академика Зуева, 1
Заказ № 94. Тираж 100 экз.