На правах рукописи

Атаева Ольга Муратовна

РАЗРАБОТКА И РЕАЛИЗАЦИЯ СЕМАНТИЧЕСКОЙ ЦИФРОВОЙ БИБЛИОТЕКИ КАК ОСНОВЫ ДЛЯ ПОСТРОЕНИЯ ПРОСТРАНСТВА НАУЧНЫХ ЗНАНИЙ

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

Автореферат

диссертации на соискание ученой степени
кандидата технических наук

Москва – 2019

Работа выполнена в Федеральном исследовательском центре «Информатика и управление» Российской академии наук

Научный руководитель: **Серебряков Владимир Алексеевич** доктор физико-математических наук, профессор, главный научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук

Официальные оппоненты: **Аветисян Арутюн Ишханович** доктор физико-математических наук, чл.-кор. РАН, профессор РАН, директор Института системного программирования им. В.П. Иванникова Российской академии наук.

Елизаров Александр Михайлович доктор физико-математических наук, профессор, профессор кафедры программной инженерии высшей школы информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета

Ведущая организация: Федеральное государственное учреждение «Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук»

Защита состоится «22» января 2020 г. в 16 часов 00 мин. на заседании диссертационного совета Д 002.073.02 при Федеральном исследовательском центре «Информатика и управление» Российской академии наук по адресу 119333, г. Москва, ул. Вавилова, д. 44, к. 2.

С диссертацией можно ознакомиться в библиотеке Федерального исследовательского центра «Информатика и управление» Российской академии наук по адресу 119333, г. Москва, ул. Вавилова, д. 44, к. 2 и на сайте www.frccsc.ru.

Автореферат разослан «_____» _____ 2019 г.

Ученый секретарь
диссертационного совета Р.В. Разумчик

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Актуальность работы связана с тем, что последние десятилетия объем информации лавинообразно увеличивается и это касается и научных областей. Продолжаются попытки построить формальные модели **научных** предметных областей, например математических. Увеличивается количество источников разнообразных данных и форматов, в которых они представлены. Резко возросло время, необходимое для поиска нужной информации и ее обзора. Главной задачей создания описания обобщенного представления научных знаний для некоторой области является помощь экспертам в организации знаний и предоставления доступа к ней. При этом средство организации знаний должно быть достаточно универсальным и не требовать глубоких технических познаний.

Говоря далее о произвольных предметных областях, имеются в виду предметные области, которые относятся к различным отраслям науки, например, такие как математика, микробиология и т.д. Главная особенность таких предметных областей заключается в том, что перечень и структура ресурсов таких областей подвержена частым изменениям.

Говорить о представлении научного знания можно, исходя из признания того положения, что знание целостно, а деление его на отдельные дисциплины лишь условно. В связи с этим появление в последние десятилетия гипертекста, метаинтерпретаций и метаязыков, для описания, в частности научного знания, определяет новый тип семантических библиотек и ее пользователей, которые не только потребляют, но и являются полноценными участниками конструирования структуры знания предоставляемого этими библиотеками, в соответствии со своим языком.

Целью диссертационной работы является задача создания такой информационной системы для библиотек, которая могла бы учитывать все разнообразие различных типов ресурсов определенной научной предметной области, которые могут в ней храниться и при этом поддерживать ее терминологическое описание. Одна из основных решаемых задач в контексте системы – это обеспечение возможности интеграции данных из источников поддерживающих семантическое описание модели данных. Фактически такая система должна представлять собой конструктор для создания цифровой библиотеки любой направленности и с адаптируемой моделью контента хранимых данных. Адаптируемая модель данных позволит описывать произвольную модель данных контента библиотеки в рамках фиксированной в терминах тезауруса предметной области.

Решаемая научная задача заключается в разработке модели информационных ресурсов и объектов, а также модели терминологического описания в виде тезауруса научных электронных семантических библиотек. Предлагаются методы семантической классификации информационных объектов на основе тезауруса, учитывающие связи между ними, что дает возможность более полно формировать картину научного знания в рассматриваемой области. Предлагаются алгоритмы интеграции данных в пространство научных знаний из различных источников данных.

Реализация поставленной цели предполагает решение следующих **подзадач**:

1) разработка онтологического представления контента библиотеки, которое позволяло бы описывать любые типы ресурсов, включаемых в библиотеку;
2) разработанная модель представления должна легко интегрироваться с любой предметной областью, представление которой ограничивается набором ее терминов в виде некоторой таксономии (линейный словарь, классификатор, тезаурус);
3) разработка расширяемой понятийной модели представления тезауруса для поддержки сложно структурированных отраслевых тезаурусов научного знания;
4) разработка информационной системы библиотеки, в основу модели данных которой положена разработанная онтологическая модель контента библиотеки;
5) представление данных разработанной информационной системы библиотеки должно быть согласовано с требованиями, предъявляемыми к данным и источникам в рамках Linked Open Data (далее LOD);
6) реализовать поддержку семантической разметки описаний контента библиотеки с помощью тезауруса предметной области;
7) информационная система библиотеки должна поддерживать для пользователей возможность определения круга своих интересов с использованием предметного тезауруса, с возможностью его расширения для терминологического описания интересующего пользователя направления.

Одна из основных целей разрабатываемого решения – это интеграция и связывание данных библиотеки с данными из различных источников. Основные задачи, решаемые на этом этапе – устранение проблем, возникающих при объединении данных из разных источников, как на уровне данных, так и на уровне схем данных. В список подзадач для реализации поставленной цели включаются следующие:

8) информационная система библиотеки должна поддерживать интеграцию модели данных с различными источниками данных из LOD;
9) предоставлять данные библиотеки в машиночитаемом формате;
10) поддерживать механизмы связывания данных библиотеки с данными из других источников.

Результаты, выносимые на защиту:

1) Разработан подход к построению обобщенной модели научной предметной области, который делает упор на выделении таких метаданных, которые позволяют проектировать конкретные структуры данных для различных научных предметных областей и выявить общие подходы к управлению этими данными и их обработке.
2) Предложена общая модель интеграции научных знаний в рамках предметной области на основе обобщенной модели описания информационных ресурсов, которая определяет возможности интеграции с

различными источниками данных, что позволяет обогащать их с использованием интерфейсов библиотеки.

3) Предложен подход к реализации семантических информационных систем, способных гибко настраиваться под запросы конкретной предметной области и формирования ее онтологии на основе высокоуровневых понятий обобщенной модели научной предметной области.

4) Усовершенствован доступ и восприятие больших и сложно структурированных объемов информации пользователем на основе тезауруса, учитывающего связи между объектами семантической библиотеки, что дает возможность более полно формировать картину научного знания в рассматриваемой области.

5) Разработана гибкая, настраиваемая модель поддержки тезауруса в семантической библиотеке, которая позволяет выявлять и фиксировать новые связи между элементами тезауруса и контентом библиотеки, позволяя фиксировать научные знания в структурированном виде.

Объектом исследования являются основные понятия научных предметных областей и их использование в электронных библиотеках.

Предметом исследования является использование семантических технологий Semantic Web для реализации научной электронной библиотеки в определенной области знания.

Методы исследования. Для решения поставленных задач в работе использовались методы системного анализа и семантического моделирования, теории графов и множеств, объектно-ориентированного проектирования и программирования, методы обработки научных текстов и методы поддержки терминологического описания научной предметной области.

Научная новизна заключается в том, что в отличие от существующих решений предложена онтология, состоящая из минимального количества классов и свойств, что позволяет ее использовать в качестве базовой онтологии для построения более сложных моделей данных различных научных предметных областей. Отличительной особенностью является то, что обобщенный подход к описанию контента, позволяет реализовать средства интеграции данных в рамках библиотеки, адаптируемые под условия любой предметной области без оглядки на ее специфику. Это позволяет формализовать процесс интеграции внешних источников в библиотеку, и благодаря гибкости адаптивной модели, позволяет интегрировать любой необходимый источник, удовлетворяющий требованиям в рамках введенных понятий. В отличие от разработанных ранее решений для семантических библиотек, обеспечивается конструирование такой библиотеки не на уровне разработки программного обеспечения, а на уровне настройки уже установленной системы, что существенно сокращает время ее построения. Оригинальность подхода заключается в том что, появляется возможность смоделировать пространство знаний с точки зрения его отображения в контексте семантической библиотеки, решающей как классические задачи сбора сохранения навигации по ее содержимому, так и по извлечению новых, неявно определенных связей между объектами, составляющими наполнение библиотеки. В работе предложен новый подход к модели тезауруса, что позволяет реализовать гибкий настраиваемый поиск,

результатом которого будет сбалансированный список объектов по предметной области, и настраивать модель данных тезауруса под специфику предметной области.

Теоретическая значимость исследования заключается в том, что полученные в диссертационной работе результаты вносят вклад в развитие теории создания нового поколения информационных систем ориентированных на научные предметные области, основанных на онтологическом моделировании и технологиях Semantic Web, извлечении знаний из источников данных на их основе и построении картины научного знания по рассматриваемой предметной области в условиях непрерывно поступающего потока информации.

Практическая значимость

1) Разработанные в работе модели, подходы и алгоритмы применены для создания программного обеспечения научных электронных библиотек для некоторой предметной области. Программное обеспечение создается на основе явного описания модели ресурсов высокоуровневых понятий научной предметной области, с использованием технологий Semantic Web.
2) Разработанные прототипы программных систем могут быть использованы для конструирования научных семантических электронных библиотек с использованием технологий Semantic Web.

Реализация и внедрение результатов работы. Разработанная в диссертации система построения семантических библиотек для построения научного пространства данных некоторой научной предметной области внедрена и используется в библиотеке по естественным наукам РАН по предметным областям микробиология и лингвистика. Также система внедрена и используется в межведомственном суперкомпьютерном центре Российской академии наук для сопровождения информационных ресурсов в рамках электронной библиотеки «Научное наследие России» в ходе проведения работ по интеграции библиотеки в Linked Open Data.

Личный вклад. Выносимые на защиту результаты получены соискателем лично. В опубликованных совместных работах постановка и исследование задач осуществлялись совместными усилиями соавторов при непосредственном участии соискателя.

Апробация работы. Основные положения диссертации изложены в 16 публикациях. По теме диссертации были сделаны сообщения и доклады на международных научно-практических конференциях, симпозиумах и форумах: Международная научная конференция «Информационные технологии и системы. Наука и практика» (Владикавказ, 2009г.), 55-ая Научная конференция МФТИ, (Долгопрудный, 2012 г.), Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL (Ярославль, 2013г.), VII Национална конференция „Образованието и изследванията в информационното общество” (София, 2014г.), XVII Всероссийская научная конференция ИПМ им. М.В.Келдыша, Научный сервис в сети Интернет труды (Новороссийск, 2015г.), XVIII Международная научная конференция «Аналитика и управление данными в областях с интенсивным использованием данных» (“Data Analytics and Management in Data Intensive Domains”) (DAMDID/RCDL'2016), (Москва, 2016г.), Общероссийский

семинар Информатика, управление и системный анализ (Москва, 2017г.), XX Всероссийская научная конференция ИПМ им. М.В.Келдыша, Научный сервис в сети Интернет труды (Новороссийск, 2018г.), Международная научно-практическая конференция Математическое образование в школе и вузе: инновации в информационном пространстве (MATHEDU, Казань, 2018 г.)

Публикации. По материалам диссертации опубликовано 16 работ, из них 5 статей в изданиях, входящих в перечень ВАК, 11 статьей в сборниках трудов конференций.

Структура и объем диссертационной работы. Диссертация состоит из введения, шести глав и заключения. Каждая глава завершается выводами. Полный объем диссертации составляет 157 страниц. Список литературы содержит 160 наименований. В диссертации 15 рисунков, 25 таблиц, приводится 1 приложение. Объем приложения составляет 7 страниц.

СОДЕРЖАНИЕ РАБОТЫ

Во введении аргументируется актуальность исследований, формулируются цель и задачи работы, перечисляются используемые методы исследования, обосновывается научная новизна и практическая значимость полученных результатов, приводятся сведения о результатах внедрения и использования.

Первая глава посвящена обзору основных концепций и инструментов, которые легли в основу работы. Проведен анализ существующих решений по семантическому представлению знаний и интеграции представленных данных. В главе осуществляется постановка задачи и описывается логическая схема исследования.

В разделе 1.1 рассматривается парадигма Semantic Web предложенная в 2006 году Тимом Бернерс-Ли. Он сформулировал основные принципы связанных данных - надстройки над существующим интернетом, которая позволяет автоматизированным системам извлекать информацию, анализировать её, устанавливать взаимосвязи и генерировать новую информацию. Для этого было предложено использовать модель представления данных RDF (Resource Description Framework), которая пригодна для машинной обработки. Структурно выражения в RDF представляют собой триплеты, состоящие из субъекта, предиката и объекта. Выражение RDF триплета означает, что отношение, указанное предикатом, связывает предметы, обозначенные как субъект и объект.

В разделе 1.2 рассматриваются онтологии, которые стали активно использоваться с появлением парадигмы Semantic Web для формализации знаний в различных предметных областях.

В подразделе 1.2.1 рассматривается онтологический подход, при котором любая информационная система на концептуальном уровне обладает общим набором понятий, которые описывают понятия любой информационной системы. Наиболее полная онтология для описания информационных систем является онтология BWW (Bunge, Wand and Weber ontology), которая фокусируется на модели представления и определяет набор понятий, их связей и характеристик, достаточных для описания структуры и поведения информационных систем.

В разделе 1.3 рассматриваются тезаурусы, которые в отличие от онтологий предназначены для терминологической поддержки описания предметной области. Для описания какой-либо предметной области всегда используется определенный набор терминов, каждый из которых обозначает или описывает какое-либо понятие или концепцию из предметной области. Совокупность терминов, описывающих предметную область, с указанием семантических отношений (связей) между ними и является тезаурусом. Такие отношения в тезаурусе всегда указывают на наличие смысловой (семантической) связи между терминами.

В разделе 1.4 приводится концептуальная модель электронных библиотек с определениями важнейших представлений об архитектуре, ресурсах и функциональности электронных библиотек, которая была определена в программном документе DELOS (Digital Library Reference Model, DLRM). В ней были определены базовые для электронной библиотеки понятия (конкретная ЭБ, система ЭБ, система управления ЭБ), выделены категории пользователей для этих понятий (разработчик, пользователь, администратор). Выделены шесть основных высокоуровневых понятий/областей: (1) контент, (2) пользователь, (3) функциональные возможности, (4) качество, (5) политики, (6) архитектура.

В разделе 1.5 приводится обзор решений для создания семантических библиотек, созданных разными исследовательскими группами.

В подразделе 1.5.1 в качестве первого примера рассматривается популярная система Greenstone3, которая свободно распространяется, и главной целью которой является создание и поддержка пользовательских коллекций цифровых ресурсов самых разнообразных типов. Система довольно легко настраивается и устанавливается. Пользователи системы могут создавать свои коллекции, включая самые разнообразные типы ресурсов и обеспечить к ним доступ, как через локальную, так и через глобальную сеть.

В подразделе 1.5.2 рассматривается семантическая библиотека JeromeDL, которая является одной из попыток объединить возможности, предлагаемые концепцией и технологиями Semantic Web, с библиотеками, ориентируясь на тесное взаимодействие с пользователями. Фактически она представляет собой интегрированное приложение для ведения цифровой библиотеки, блогов и сервиса для закладок. В рамках цифровой библиотеки поддерживаются авторитетные файлы (для авторов, редакторов, издательств), таксономии, используемые для классификации по темам, тезаурус WordNet для ключевых слов. Каждый ресурс описывается тремя типами метаданных: структурными, библиографическими и социальными. Каждый тип метаданных поддерживается соответствующими сервисами. Пользователю предоставляется комбинированное представление на основе этих метаданных. Основные модели для описания ресурсов, пользователей и их взаимодействия - это библиографическая онтология MarcOnt, онтология FOAF, модель знаний SKOS для описания таксономий.

В подразделе 1.5.3 рассматривается одна из глобальных реализованных цифровых библиотек проект Europeana, который интегрирует данные из институтов культурного наследия Европы. Многоуровневая организация провайдеров контента предназначена для автоматической оценки контента на соответствие модели данных EDM (Europeana Data Model), которая была

разработана в рамках проекта. В рамках этой модели данных определены наборы классов и свойств для описания объектов культурного наследия. Одно из преимуществ EDM - это возможность соблюдения принципов связанных данных при описании ресурсов.

В разделе 1.6 осуществляется постановка задачи и приводится логическая схема исследования. Сформулирована задача создания такой информационной системы для библиотек, которая могла бы учитывать все разнообразие различных типов ресурсов определенной научной предметной области, которые могут в ней храниться и при этом поддерживать ее терминологическое описание. Одна из основных решаемых задач в контексте системы – это обеспечение возможности интеграции данных из источников поддерживающих семантическое описание модели данных. Фактически такая система должна представлять собой конструктор для создания цифровой библиотеки любой направленности и с адаптируемой моделью контента хранимых данных. Адаптируемая модель данных позволит описывать произвольную модель данных контента библиотеки в рамках фиксированной в терминах тезауруса предметной области.

Исходя из постановки задачи, формулируется структурная схема исследования.

Вторая глава посвящена рассмотрению концепции, на которых базируется определение семантической библиотеки для научной предметной области. Рассматривается роль информационной системы в процессе конструирования такой библиотеки.

В разделе 2.1 рассматриваются последовательные этапы эволюции библиотек: электронные, цифровые и семантические библиотеки.

Приводится формальное определение электронных библиотек. Ее контент фактически представляет собой множество библиографических записей объектов реальной классической библиотеки. В таком представлении, например, встретив упоминание персоны в одном месте, невозможно точно установить соответствие с ее упоминанием в другом месте.

Далее приводится формальное определение цифровых библиотек. Ее контент фактически представляет собой множество мультимедийных объектов.

Также приводится формальное определение семантических цифровых библиотек, которые являются следующим этапом в эволюции библиотек и обязаны своей популярностью семантическим технологиям, которые в значительной степени повлияли на переосмысление понятия контента библиотеки и послужили толчком для расширения и улучшения функциональности библиотек. В таких библиотеках данные лучше структурированы, выделены связи между ними, улучшается поиск, появляется возможность интеграции данных различных типов. Обеспечивается интероперабельность с другими системами, не обязательно являющимися библиотеками, так как основной задачей семантических технологий является предоставление метаданных в машиночитаемом формате.

В разделе 2.2 приводятся основные свойства семантических библиотек. Главной возможностью таких библиотек является структурирование их разнообразного контента и возможность связывания данных из разных

источников, что в свою очередь, несомненно, отражается на качестве данных контента.

В разделе 2.3 проводится выделение модели контента семантической библиотеки и ее основные характеристик, Было отделено определяющее понятие контента семантической библиотеки от понятия реализующей библиотеку информационной системы. Такой подход позволяет наращивать функциональность системы, добавлять новые подсистемы или изменять уже имеющиеся при неизменных остальных частях.

В разделе 2.4 приводится построение обобщенной модели научной предметной области, представляется подход, который делает упор на выделении таких метаданных, которые позволяют проектировать конкретные структуры данных для различных научных предметных областей и выявить общие подходы к управлению этими данными и их обработке. Это позволит построить общую модель научных знаний, в рамках которой можно проводить интеграцию данных из различных источников.

Третья глава посвящена рассмотрению основных понятий информационной модели, которые позволяют выстроить систему для конструирования семантической библиотеки таким образом, чтобы выполнялись основные требования, предъявляемые к таким библиотекам.

С одной стороны понятийная структура семантических библиотек не является устоявшейся и разнится в зависимости от конкретной реализации. С другой стороны, эффективность исследований в этой области зависит от стандартизации и формализации собственно описаний ресурсов таких библиотек и процессов их представления.

Выделяя явным образом набор понятий, с помощью которых можно описать содержимое библиотеки, явно выделяемое в стандарте DELOS, даются определения для построения формальной модели типов ресурсов, которые лежат в основе ее построения. Вводимые понятия в дальнейшем помогут формировать понятийную основу конкретной предметной области для описания ее пространства научных знаний.

Фактически, понятия делятся на три категории: первая включает определения понятий контента семантической библиотеки и вторая категория относится к определению понятий необходимых для поддержки терминов в тезаурусе предметной области и третья включает определения, необходимые для определения процессов интеграции контента этих ресурсов. На основе этих определений описываются основные процессы такие, как, например, интеграция данных из разных источников, категоризация/классификация, отображение разных моделей данных источников на заданную предметную область, построение классов эквивалентности и т.д.

В разделе 3.1 приведены основные свойства разрабатываемой модели семантической библиотеки научного пространства знаний.

В разделе 3.2 описана модель контента библиотеки и введены основные понятия: информационный объект, информационный ресурс, атрибут, набор атрибутов, свойства и связи между ними. По своим функциям атрибуты можно делить на следующие пересекающиеся виды атрибутов: идентифицирующие, озаглавливающие, обязательные, классифицирующие, поисковые, описательные. В процессе обработки данных каждая задача в системе использует определенный

тип атрибутов. Например, классифицирующие атрибуты обеспечивают поддержку задач классификации объектов и сохранения ее результатов. В соответствии с типом значений атрибуты могут быть однозначными, многозначными, т.е. имеющими множество однотипных значений, которые могут составлять мультимножество, множество, список, массив.

В разделе 3.3 описана модель тезауруса. Вводятся основные понятия концептов, терминов, дескрипторов, недескрипторов, связи между концептами, связи между концептами и контентом библиотеки. Вводятся понятия атрибутов тезауруса и набора атрибутов тезауруса. Эти понятия используются для создания расширенного описания структуры концепта тезауруса и определения дополнительных связей с контентом библиотеки.

В разделе 3.4 описана модель интеграции данных библиотеки с внешними источниками данных. Исходя из основных понятий введенных ранее, модель контента библиотеки G представляет собой множество ресурсов $R = \{r_j\}$, множество атрибутов $A = \{a_i\}$ и для каждого ресурса определен набор атрибутов $N(r) \subset A$, то есть $r_j(a_1,.., a_n)$, $a_n \in N(r)$. В каждый набор атрибутов входят так называемые идентифицирующие атрибуты, обозначим их как $I(r) \subset N(r) \subset A$, для однозначной идентификации информационных объектов этого ресурса.

Формально подсистема интеграции I_T представляется тройкой $<G, \{S_i\}, \{M_i\}>$, где G – предварительно определенная модель контента, состоящая из множества ресурсов R и их описаний в виде набора атрибутов N(r), S_i – схема i-го источника подключенного к системе, M_i – отображение i-го источника, $1 \leq i \leq n$, где n количество источников данных.

В подразделе 3.4.1 вводятся операции, которые делают подсистему интеграции настраиваемой под любой источник данных с возможностью обогащения уже имеющихся в системе данных. Динамическое доопределение модели контента библиотеки фактически включает в себя этап анализа ресурсов интегрируемого источника данных и расширение или уточнение исходной модели контента путем расширения множества информационных ресурсов или множества атрибутов. Возможность выполнения этих операций обеспечивается благодаря принятой адаптивной модели данных системы.

В подразделе 3.4.2 определяется процесс построения отображения, который можно разделить на несколько основных этапов: подключение источника данных, определение типов ресурсов библиотеки, определение отображения атрибутов. Благодаря такому построению отображения получается набор правил, по которым можно представить каждый найденный объект в источнике в рамках понятий библиотеки и соответственно позволить его сохранить полностью в локальном хранилище по требованию пользователя, либо просто сохранить связь между найденным объектом в источнике и объектом в библиотеке.

В подразделе 3.4.3 определяется процесс построения поисковых запросов по источникам данных после построения отображения. То есть, любой экземпляр ресурса источника данных, для которого построено отображение, может являться ответом на запрос пользователя.

Подключая источники данных из LOD для интеграции данных на уровне ресурсов источника и библиотеки, используется связь или отображение, которая указывает, что два разных класса могут иметь одинаковых представителей. Это

отображение может указывать на класс в источнике данных LOD, который является источником дополнительной информации о ресурсе – субъекте или на эквивалентный ему класс, возможно с разной степенью детализации описания объектов. Фактически предполагается, что онтология источника данных частично совместима со структурой ресурсов библиотеки. Это означает, что хотя бы один ресурс онтологии может быть транслирован в некоторый класс в онтологии источника данных. Требуется лишь минимальное частичное соответствие ресурсу библиотеки. Для однозначной идентификации экземпляров соответствующего класса из источника данных, отображаться должны как минимум идентифицирующие атрибуты, определенные для информационного ресурса в системе.

В разделе 3.5 описаны семантически значимые связи, определенные в рамках обобщенной модели научной предметной области.

Самый простой набор правил выявления неявных связей определен в самом стандарте онтологического представления информации. Помимо них можно использовать возможность явно определять правила, согласно которым должны получаться логические выводы. Синтаксис правил довольно прост и состоит из двух частей: первая часть определяет условие, при выполнении которого во второй части определяется вывод.

В четвертой главе дано описание онтологии научного пространства знаний, которое может быть представлено с точки зрения двух ортогональных подходов:

1. вводятся термины, характерные для рассматриваемой научной предметной области, соединенные различными связями как иерархическими, так и горизонтальными;
2. вводится набор определений, который на более абстрактном уровне описывает множества объектов научной предметной области, фактически задавая структуру их описания и отношений между ними.

В различных исследованиях в обоих случаях говорят или о построении тезауруса предметной области, или о построении онтологии предметной области. Но это два совершенно разных подхода к описанию предметной области, которые не являются при этом взаимоисключающими. Такой подход, с одной стороны, позволяет отдельно сконцентрироваться только на типах информационных ресурсов библиотеки, которые являются ресурсами пространства знаний, и описать основные понятия, характерные для этой предметной области. С другой стороны, говоря о тезаурусе, будем иметь в виду набор понятий и терминов, которые обеспечивают терминологическую поддержку понятий онтологии предметной области.

В разделе 4.1 приводится алгоритм построения онтологии конкретной предметной области, описанной в терминах введенной выше онтологии семантической библиотеки. При этом если новые введенные понятия являются на первом уровне экземплярами обозначенных ресурсов, то при наполнении библиотеки мы используем их в качестве классов для описания данных. Рассмотрение экземпляров в качестве классов называют метамоделированием. И хотя даже прямая семантика языка онтологий OWL2, используемого для описания онтологий, не позволяет такого метамоделирования, это ограничение в языке обходится с помощью синтаксического трюка известного под название

punning. Это означает, что когда идентификатор экземпляра встречается в аксиоме класса, то он рассматривается как класс, а когда этот же идентификатор встречается в отдельном утверждении, то рассматривается как экземпляр.

Итак, выполняя описание конкретной предметной области в терминах разработанной онтологии семантической библиотеки, мы фактически конструируем трехуровневую онтологию, в которой экземпляры первого уровня - это высокоуровневые понятия, на втором уровне мы описываем понятия конкретной предметной области как экземпляры в терминах первого уровня и используем их как определения классов на третьем уровне при заполнении онтологии данными.

В разделе 4.2 приводятся примеры использования правил вывода новых знаний. Применение правил имеет ряд преимуществ. Рассмотренные выше классы онтологии и отношения между ними представляют собой факты или знания о предметной области. Занесение всех фактов и знаний о классах и их экземплярах в пространство знаний, смоделированном с помощью онтологии, может потребовать достаточно много времени. Если в онтологию занести лишь первичные факты или знания между классами и их экземплярами, то часть вторичных фактов или знаний мы можем вывести с помощью правил, описывающих вывод вторичных на основе первичных. Также правила позволяют устранить некоторые ограничения выразительности онтологии и позволяют выводить наличие отношений между экземплярами, то есть использовать бинарный предикат, задающий отношение между объектами, тогда как в OWL мы можем определять только унарный предикат, определяющий класс. С помощью логического вывода можно автоматически классифицировать экземпляры, представленные информационными объектами, по классам, представленным информационными ресурсами, на основе их атрибутов.

В пятой главе рассматривается формальное описание системы, определяющее ее цели, функции, внешне видимые свойства, и интерфейсы. Оно также включает описание компонентов системы и их отношений, наряду с принципами, управляющими ее дизайном, функционированием и возможным последующим развитием. Это описание включает программные подсистемы, визуализированные свойства этих подсистем, отношения между подсистемами и ограничения на их использования. При этом каждая подсистема может состоять из нескольких уровней абстракции, и каждый уровень может иметь свою архитектуру.

В разделе 5.1 приведена основная функциональность информационной системы LibMeta, разработанной на основе разработанного метода построения семантической библиотеки для некоторой области научного пространства. Основная функциональность делится на множество функций доступных для всех публичных пользователей и множество подмножеств функций авторизированного пользователя доступных ему частично или полностью в зависимости от его роли в системе.

В разделе 5.2 описана подсистема описания контента информационной системы, которая отвечает за универсальность определения контента системы. Эта подсистема обеспечивает структурированное описание контента и обеспечивает поддержку его адаптируемости. А также обеспечивает описание

конкретных ресурсов и их объектов в виде RDF троек и предоставления SPARQL точки доступа для публикации данных в LOD.

На рисунке 1 приведены основные понятия, используемые для конструирования описания предметной области в рамках этой подсистемы.

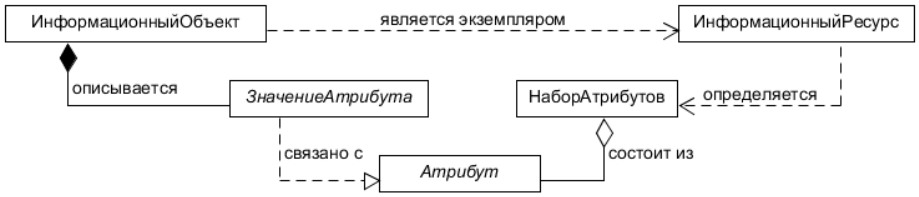

Рис. 1. Основные понятия, используемые для конструирования описания предметной области

В разделе 5.3 описана подсистема управления тезаурусом, основные функций которой позволяют определить также совокупность словарей и классификаторов, являющихся частью тезауруса, а также разнообразные связи между их понятиями. В рамках этой подсистемы для построения базовой версии тезауруса используются следующие понятия иерархическая связь, горизонтальная связь, термин, тезаурус, концепт, тематическая группа, термины, дескриптор (или предпочитаемый термин понятия), аскриптор (множество терминов, являющихся синонимами дескриптора). Эти понятия и связи между ними приведены на рисунке 2.

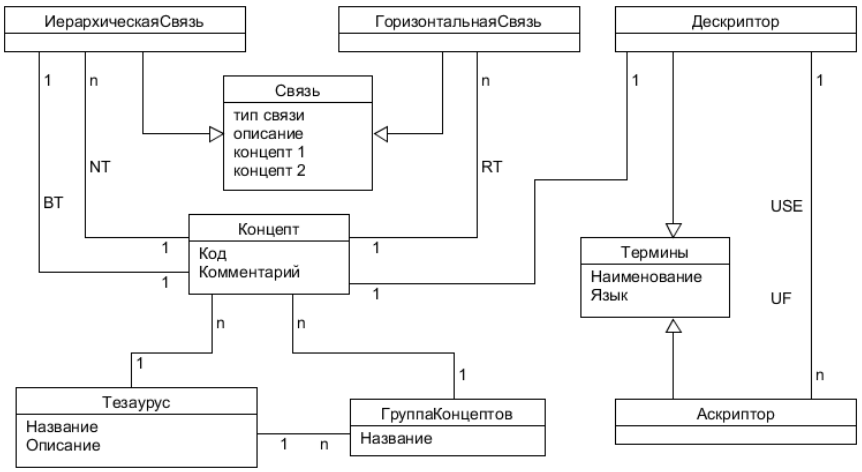

Рис. 2. Основные понятия, используемые для конструирования тезауруса предметной области

При построении тезауруса также доступна возможность доопределить атрибуты, которые позволяют расширить определение концепта и включать в его определение также информационные объекты и доопределить связи между концептами.

В разделе 5.4 описана подсистема поддержки коллекций, которая определяется на основе некоторой таксономии с указанием коллекционируемых типов ресурсов. Коллекция может объединять информационные объекты различных информационных ресурсов. На основе одной той же таксономии

можно определять несколько коллекций. Такой подход оказывается чрезвычайно полезным для создания раздельных пользовательских коллекций.

Явное описание коллекций в рамках этой подсистемы, позволяет поддержать механизм так называемой гибкой классификации информационных объектов определяемых ресурсов. Словом «гибкая» подчеркивается, что возможность организации/классификации объектов в соответствии с тезаурусом/тематическим словарем/тематическим классификатором может настраиваться на любом этапе жизнедеятельности системы, никак не влияя на решения, принятые на этапе моделирования предметной области.

В разделе 5.5 описана подсистема автоматизированной обработки и представления данных. Эта подсистема позволяет автоматизировать операции создания и редактирования объектов. Входящие данные последовательно проверяются на соответствие модели данных, после чего загружаются в систему.

Для предотвращения дублирования информационных объектов загрузчик использует алгоритм, который отвечает за поиск в репозитории схожих объектов, их ранжирование, а по возможности также автоматическое принятие решения о слиянии входящего и имеющегося объектов. В случае невозможности принятия такого решения без участия пользователя, интегратор передаёт сообщение для пользователя.

Входной формат данных представляет собой сокращенный вариант RDF/XML синтаксиса.

В разделе 5.6 описана подсистема реализации задач интеграции данных из источников LOD, в рамках которой каждому источнику данных ставятся в соответствие информационные ресурсы системы, и устанавливается соотношение набора атрибутов ресурса со свойствами ресурса из источника данных. Это позволяет нам генерировать SPARQL запросы к источникам данных для извлечения конкретной информации. При этом пользователь оперирует привычными формами поиска, избегая необходимости написания самих запросов.

В разделе 5.7 описана подсистема поддержки пользователей LibMeta. Важной составляющей любой информационной системы являются ее пользователи. Для каждого пользователя уровень доступа определяется его ролью, определяющей набор прав доступа для работы с информационными ресурсами и объектами. Для каждого пользователя системы определяется его область интересов, в описании которой может быть задействован тезаурус предметной области контента библиотеки, а также список пользователей со сходным кругом интересов. На рисунке 3 приведены основные понятия, используемые в рамках этой подсистемы.

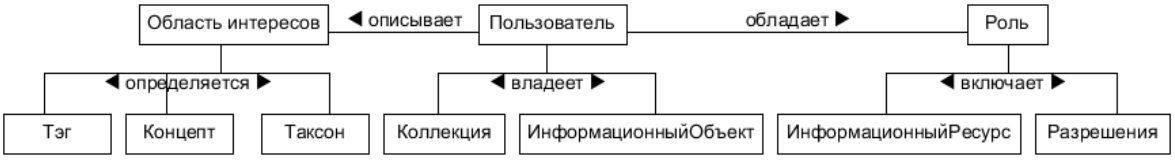

Рис. 3. Основные понятия, используемые для описания уровней доступа пользователя к информационным ресурсам и объектам предметной области в рамках библиотеки

В разделе 5.8 описана подсистема поддержки микротезауруса пользователя. Основная функциональность подсистемы связана с определением структуры микротезауруса, которая полностью совпадает с основным тезаурусом, определенным в системе в рамках некоторой предметной области. Особенностью микротезауруса является то, что главным узловым элементом выбирается концепт основного тезауруса и пользователь может развивать ветку тезауруса на его основе. Микротезаурусы пользователей – экспертов могут помочь редакторам основного тезауруса принять решение о расширении основного тезауруса на их основе.

В разделе 5.9 описана рекомендательная подсистема. Рассматриваемая подсистема рекомендаций основывается как на анализе текстовой информации из метаданных информационных объектов, так и, при наличии, текстов являющихся содержимым этих объектов. Так же важную роль играют связи, которые связывают концепты тезауруса и информационные объекты. Функциональность рекомендательной подсистемы будет доступна пользователям при описании своей области интересов. Помимо рекомендаций по области интересов поэтапно проводится анализ информационных объектов связанных с этими понятиями. При наличии связанных с ними объектов они объединяются в категории и предоставляются пользователю в качестве рекомендации.

В шестой главе рассматривается программная реализация семантической библиотеки LibMeta. Настоящая глава посвящена описанию и анализу характеристик прототипа этой системы, созданного автором на основе разработанной в диссертации архитектуры. Описываются возможности прототипа и приводятся результаты его исследования на соответствие предъявляемым к системе требованиям, которые сформулированы при постановке задачи.

В разделе 6.1 рассматриваются особенности программной реализации. Основной код прототипа системы написан на языке Groovy с использованием фреймворка Grails. Фреймворк Grails распространяется в открытых исходных кодах по лицензии Apache License 2.0. В качестве шаблона проектирования программного комплекса в Grails используется широко распространенный шаблон схема «модель-представление-поведение» (Model-View-Controller, MVC). Использование этого шаблона проектирования облегчает понимание, написание, модификацию и диагностику программного кода за счет разделения трех основных частей программного комплекса – модели данных, представления данных и контроллера данных, который является связующим звеном между пользователем и системой.

В разделе 6.2 рассматриваются примеры практической апробации системы LibMeta. Описание тезауруса каждой предметной области в терминах понятий базовой версии онтологии расширяется дополнительно с помощью понятий расширенной модели для возможности расширения структуры статьи понятия конкретного тезауруса.

В подразделе 6.2.1 рассматривается в качестве примера реализации семантической библиотеки, на основе изложенной в работе модели предметная область обыкновенных дифференциальных уравнений (ОДУ). На основе разработанной модели было выполнено конструирование библиотеки для этой

области. В качестве тезауруса использован тезаурус ОДУ, разработанный коллективом специалистов в этой области.

В подразделе 6.2.2 рассматривается в качестве примера предметная область задач математической физики. Так как область уравнений математической физики, куда входят уравнения в частных производных, как предметная область, включает в себя необъятное количество материала, в тезаурусе ограничиваются вопросами определения терминологии для **идентификации физических процессов**, как основы для математических моделей и **для уравнений в частных производных** с примерами **из уравнений смешанного типа**.

В подразделе 6.2.3 рассматривается в качестве примера предметная область «Микробиология и физиология растений». Структура тезауруса данной предметной области не содержит глубоких иерархий, но содержит множество горизонтальных связей между понятиями. Особенность понятий рассматриваемого тезауруса такова, что они имеют мультидисциплинарный характер. Поэтому некоторые из них содержат указание на смежную область науки или явную ссылку на тезаурус смежной предметной области. Также для каждого понятия указывается соответствующий код УДК и/или ББК. Это позволяет уточнять семантику связанных статей и использовать пристатейные ключевые слова и термины тезауруса как ключевые слова соответствующих рубрик классификаторов УДК и ББК.

В подразделе 6.2.4 рассматривается электронная версия советской математической энциклопедии 1978 года, которая используется в качестве тезауруса предметной области. Структура понятий математической энциклопедии не обладает иерархией как таковой, но благодаря использованию связанных с понятиями кодов MSC мы смогли выделять тематически связанные термины отдельных разделов математики. Из статей были выделены упоминаемые персоны и проставлены связи между понятиями и персонами. Были отдельно проиндексированы формулы, и каждому понятию при возможности был сопоставлен набор соответствующих формул.

В подразделе 6.2.5 обобщен опыт подключения реляционных источников данных в вышеописанных примерах. И хотя рассматриваемая информационная система включает в себя функциональность подключения источников удовлетворяющим требованиям сообщества LOD, реляционные источники подключались с помощью использования сторонних инструментов, а именно платформы D2RQ. В составе этой платформы имеются инструменты для автоматизированного отображения реляционной базы в RDF граф.

В разделе 6.3 рассматривается организация взаимодействия системы с внешними программами, не являющимися web-браузерами. Необходимо обеспечивать программных клиентов возможностью оперировать данными, находящимися в системе, для этого был реализован общедоступный программный интерфейс (API), который позволяет удаленно обращаться к функциям приложения и выполнять какие-либо действия в нем. Такой прикладной интерфейс был реализован на технологии REST (Representational State Transfer – передача представления состояния), ставшей популярной в последние несколько лет, и прекрасно подходящей для нашего приложения. Удобством использования REST является возможность проверки получаемых

данных в обычном браузере, либо использованием стандартных приложений для выполнения HTTP запросов

Возьмем в качестве примера информационный объект, который является отдельным элементом контента библиотеки, который можно извлечь, изменить или удалить. На рисунке 4 представлены ключевые понятия технологии REST: информационный объект, который является экземпляром информационного ресурса предметной области, представление состояния информационного объекта – представляет собой описание состояния объекта в терминах онтологии предметной области в RDF/XML – разметке, возвращаемого клиенту.

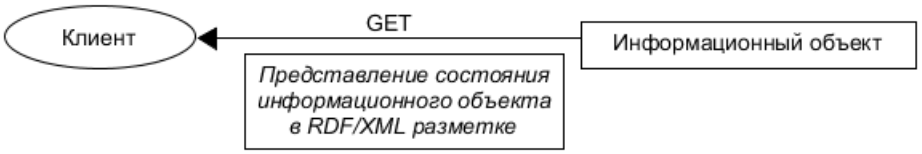

Рисунок 4. Ключевые понятия технологии REST

Как видно главное преимущество этой концепции заключается в ее простоте и в том, что протокол HTTP предоставляет практически готовую реализацию (сама по себе технология REST фактически является шаблоном проектирования).

В разделе 6.4 кратко сформулированы дальнейшие перспективы развития системы, основное направление работ связано с междисциплинарными областями знаний и формированием семантических связей между разными предметными областями и навигацией по связанным ресурсам

В заключении формулируются основные результаты и выводы диссертационной работы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

В настоящей работе представлено описание подходов и методов для построения семантической библиотеки в рамках научной предметной области. Теоретической основой работы послужил подход, основанный на онтологиях. Представлено описание общей онтологии научного пространства знаний. Разрабатываемая версия семантической библиотеки, предлагаемая автором, содержит такие модули как модуль построения онтологии ресурсов предметной области, модуль построения онтологии тезауруса предметной области, модуль загрузки информации, модуль построения интеграции и построения запросов к источникам данных, рекомендательный модуль. Разработан способ интеграции данных библиотеки с внешними источниками удовлетворяющих требованиям LOD, который является практическим воплощением парадигмы Semantic Web.

Разработанный онтологический подход к описанию научной предметной области обеспечивает выразительность достаточную для его использования при реализации основных функций семантической библиотеки. Разработанная система прошла апробацию, на которую получены акты внедрения, результаты которых подтвердили качество разработанных подходов, представленных в данной работе.

Основными научными и практическими результатами диссертационной работы являются:

- Структура и формат высокоуровневой онтологии для научного пространства знаний и представления данных в нем для семантической библиотеки.
- Разработаны и реализованы способы семантической классификации информационных объектов на основе тезауруса, учитывающие гибкие (настраиваемые) связи между ними, что дает возможность более полно формировать картину научного знания в рассматриваемой области.
- Разработаны алгоритмы решения задач поиска, автоматической категоризации, формирования рекомендаций, использующие описание модели информационных ресурсов и возможности интеграции с различными источниками данных, что позволяет обогащать данные с использованием интерфейсов библиотеки.
- Предлагается решение задачи автоматической настройки интерфейсов системы под различное семантическое описание предметной области.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

1. **А.Б. Антопольский, Атаева О.М, Серебряков В.А Среда интеграции данных научных библиотек, архивов и музеев «LibMeta» // «Информационные Ресурсы России» №5, 2012**
2. Атаева О.М., Серебряков В.А. Подход к созданию персональной семантической электронной библиотеки // Труды XXII Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL’2013. Ярославль.
3. Атаева О.М., Серебряков В.А. Персональная цифровая библиотека LibMeta как среда интеграции связанных открытых данных // Труды конференции VII Национална конференция „Образованието и изследванията в информационното общество” София, 2014
4. Атаева О.М., Кулагин М.В., Серебряков В.А. Основные понятия для построения формальной модели семантических библиотек и описания процессов интеграции в ней // В сборнике: Научный сервис в сети Интернет труды XVII Всероссийской научной конференции. ИПМ им. М.В.Келдыша. 2015. С. 8-15.
5. Малахов Д.А., Сидоренко Ю.А., Атаева О.М., Серебряков В.А. Семантический поиск как средство взаимодействия с электронной библиотекой // XVIII Международная научная конференция «Аналитика и управление данными в областях с интенсивным использованием данных» (“Data Analytics and Management in Data Intensive Domains”) (DAMDID/RCDL'2016), Москва, Россия, 11-14 октября 2016
6. **Серебряков В.А., Атаева О.М. Основные понятия формальной модели семантических библиотек и формализация процессов интеграции в ней // Программные продукты и системы. 2015. № 4. С. 180-187.**

7. Серебряков В.А., Атаева О.М. Информационная модель открытой персональной семантической библиотеки LibMeta // в сборнике Научный сервис в сети Интернет, с. 304-313
8. Malakhov D., Sidorenko Y., Ataeva O., Serebryakov V. Semantic Search in a Personal Digital Library // в сборнике Data Analytics and Management in Data Intensive Domains: XVIII International Conference, DAMDID/RCDL 2016, Ershovo, Moscow, Russia, October 11 -14, 2016, Revised Selected Papers, издательство Springer publishing co (11 west 42nd street, New York, USA, NY,10036), том 706, с. 11-14
9. Малахов Д.А., Сидоренко Ю.А., Атаева О.М., Серебряков В.А. Семантический поиск как средство взаимодействия с электронной библиотекой // в сборнике Аналитика и управление данными в областях с интенсивным использованием данных, место издания Торус пресс Москва, с. 148-154
10. **О. М. Атаева, В. А. Серебряков Персональная открытая семантическая цифровая библиотека LibMeta. Конструирование контента. Интеграция с источниками LOD // Информ. и её примен., 11:2 (2017), 85–100**
11. **Атаева О.М. Информационная модель семантической библиотеки LibMeta // Программные продукты и системы. 2016. № 4. С. 36-44.**
12. Атаева О.М., Серебряков В.А., Тучкова Н.П. Цифровая библиотека по обыкновенным дифференциальным уравнениям на основе LibMeta // Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (18-23 сентября 2017г., г.Новороссийск). — М.:ИПМ им. М.В.Келдыша, 2017. — С. 21-33.
13. Malakhov D., Sydorenko Y., Ataeva O., Serebryakov V. Semantic search in personal digital library. // Communication in Computer and Information Science. 2017. 706. 18-30. (SCOPUS)
14. **Атаева О. М., Серебряков В. А. Онтология цифровой семантической библиотеки LibMeta //Информатика и её применения. – 2018. – Т. 12. – С. 2-10.**
15. Атаева О.М., Серебряков В.А., Тучкова Н.П. Подходы к организации математических знаний при формировании предметных тезаурусов различных разделов математики // Научный сервис в сети Интернет: труды XX Всероссийской научной конференции (17-22 сентября 2018 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2018. — С. 42-54.
16. Атаева О.М., Серебряков В.А., Тучкова Н.П. Организация пространства научных знаний в области математики на примере использования тезауруса обыкновенных дифференциальных уравнений // Математическое образование в школе и вузе: инновации в информационном пространстве (MATHEDU' 2018) (Казань, 17-21 октября 2018 г.) С. 48-52