Вычислительный центр им. А.А. Дородницына Российской академии наук
Федерального исследовательского центра «Информатика и управление»
Российской академии наук (ВЦ РАН ФИЦ ИУ РАН)

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»

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель:
д. ф.-м. н., профессор
Серебряков В.А.

Москва
2019

Оглавление

Введение

До недавнего времени цифровые библиотеки воспринимались обычными пользователями как электронные версии каталогов традиционных библиотек, которые содержат описания физических объектов библиотеки (как правило, книг или других печатных изданий). Определение тематики, содержания и структуры объектов рассматриваются и воспринимались как дополнительные, но необязательные функции таких библиотек. Развитие интернета и семантических технологий вносит свои коррективы и позволяет шире взглянуть на понятие цифровых библиотек и обобщить накопившийся опыт реализации информационных систем в разных областях знаний для формирования нового типа библиотек.

Само понятие библиотеки в контексте стремительного развития интернета приобретает совершенно другой смысл и обозначает активное вовлечение пользователя в процессы, предлагаемые библиотеками. Такая библиотека предполагает участие пользователей в процессе создания, поиска и классификации того контента библиотеки, который необходим этому конкретному пользователю [140, 141, 159].

Развитие современных технологий, ко всему прочему, подталкивает к переопределению понятия *контента библиотеки*, в качестве которого не обязательно могут выступать традиционные описания печатных изданий, но и любые другие типы объектов [69, 132, 133]. Например, в качестве контента могут использоваться мультимедийные объекты: видео, звук, фотографии, музейные экспонаты, коллекции минералов, архивные материалы и многое другое. Так, например, электронная библиотека «Научное Наследие России» [69, 133], заявленная как проект по созданию библиотеки полнотекстовых научных трудов известных российских и зарубежных ученых и исследователей, включает в себя также описания музейных экспонатов, расширяя традиционные типы хранимых ресурсов классической библиотеки.

При этом необходимо четко понимать, что контент цифровых библиотек и физические объекты могут быть связаны различными способами. Это происходит из-за того, что физически объект существует в реальном мире в одном экземпляре, но в цифровой библиотеке используется лишь его описание. При этом описаний может быть несколько, они могут быть различны по структуре и смыслу, и каждое описание, ссылаясь на реальный объект, имеет собственный уникальный идентификатор, который позволяет идентифицировать конкретное описание объекта со ссылкой на него в реальном мире. Фактически цифровой ресурс может определяться как конгломерат разных описаний одного реального ресурса, представляя его общую объединенную модель.

В этом смысле интернет также может рассматриваться как библиотека, стихийно наполняемая его пользователями без видимого порядка и структуры. Каждая страница имеет собственный идентификатор в виде URL и относится к некоторому объекту реального мира, и таких страниц может быть тысячи. Чаще всего автоматическая обработка таких страниц для выявления ее взаимосвязей и полезной информации является сложной задачей и предполагает значительные затраты усилий на извлечение реально нужной информации по узкой предметной области.

Говоря в этой работе о библиотеках, имеются в виду информационные системы, которые обеспечивают основную функциональность для работы с библиотечными ресурсами, которые не ограничиваются теперь только библиографическими записями и их электронными представлениями, но также выводят на передний план семантику этих ресурсов в рамках некоторой научной области.

Вопросами семантической организации знаний занимались различные исследователи с древнейших времен. Эти исследования восходят к древнегреческим и римским философам таким, как Аристотель, Платон, Феофраст и Плиний Старший. Их идеи развивались более поздними авторами такими, как Томас Аквинский, Августин Бегемот, Уильям Оккамский, Андреа Чезальпино, Карл Линней, Рене Декарт, Джон Локк, Иммануил Кант, Джеймс

Фредерик, Чарльз Амми Каттер, Мелвил Дьюи и Шияли Раманрита Ранганатан [72, 97, 98, 99, 100]. Несмотря на их разногласия в некоторых вопросах эти философы обеспечили эффективную основу для глубокого понимания организации знаний, что нашло свое отражение в работах по формализации знаний современных исследователей [72, 73, 74, 75, 76].

Для определения семантики библиотечных ресурсов разработаны различные виды классификации – отраслевые рубрикаторы, которые позволяют более детально определить тематическую направленность ресурсов [61]. Для этого используют различные классификаторы, которые отличаются друг от друга охватом предметных областей и степенью гранулярности при классификации этих областей. Для этих целей может использоваться один из широко распространенных классификаторов, например, таких, как УДК (универсальная десятичная классификация), ББК (библиотечно-библиографическая классификация), ГРНТИ (государственный рубрикатор научно-технической информации). Эти классификаторы охватывают почти все области научного знания и перечень понятий, характерных для этих областей. Обычно эти понятия носят довольно общий характер и не отражают разнообразие направлений в каждой отдельной области научного знания.

Специализированные по конкретным областям библиотеки используют обычно свои классификаторы для систематизации своих ресурсов. Такой подход обеспечивает более детальный анализ содержания документов и соотношение смысловых понятий содержимого библиотеки с определенным направлением специализированной области знания. К таким классификаторам можно, например, отнести MSC (Mathematics Subject Classification), который используется для классификации разделов математики [128, 129].

Но зачастую этих средств описания семантики недостаточно, и со временем появляются новые требования к описанию ресурсов библиотек, что приводит как к усложнению самих описаний, так и требует значительных затрат на внедрение новых способов описаний, соответствующих текущим потребностям. Увеличивающийся поток поступающих объектов практически невозможно

обработать вручную, поэтому требуются новые методы обработки и анализа поступающих данных.

Накопленные в библиотеках данные стали доступны широкому кругу пользователей через сеть, удовлетворяя *информационные потребности* которых, функциональность цифровых библиотек становится все разнообразней. В решении задач осмысленного представления контента цифровых библиотек ключевую роль стали играть *онтологии*, позволяя представлять концептуальные модели для описания самого контента этих библиотек, основываясь на ранее разработанных форматах описания, например, таких, как MARC. Такие онтологии получили название *библиографических онтологий*, дополняя семантикой эти форматы. Фактически в них фиксируются ключевые понятия объектов, составляющих наполнение библиотеки и связи между ними. Этих понятий достаточно для описания обычной классической цифровой библиотеки для любой предметной области, в которой, как было сказано выше, представлена информация о различных печатных изданиях и, возможно, их электронные версии. Но развитие семантических библиотек и технологий способствует расширению модели, определяющей наполнение библиотеки, и этого становится недостаточно.

Одновременно с расширением модели библиотечного наполнения возникает необходимость ограничения его в рамках некоторой *предметной области*. Для этого вводится *набор терминов*, используемых для описания этой предметной области. Чаще всего эти термины организованы в виде некоторого *тезауруса* с поддержкой разнообразных связей между ними. В дальнейшем мы будем называть наполнение библиотеки с такой терминологической поддержкой некоторой предметной области *контентом семантической цифровой библиотеки* или просто *контентом*.

В фокусе предлагаемой работы будут предметные области, связанные с наукой и их особенности, Будет сделана попытка выделения общих концепций для их формальных описаний в базе знаний. Особенность этих областей заключается в том, что структура данных подвержена частым изменениям [65, 66,

141, 160]. Будем говорить об обобщенной модели научной предметной области и ее особенностях, реализациях в поисковых системах и отличий от классических подходов к поиску информации в научных массивах данных.

Современные семантические библиотеки [3, 50, 51, 52, 53, 123] предоставляют для своих пользователей большой арсенал возможностей для удовлетворения их *информационных потребностей*. Это разнообразные *средства поиска*: *атрибутный поиск*, *полнотекстовый поиск*, *поиск по коллекциям* на основе тематических классификаторов, *поиск по разнообразным типам ресурсов*, включенных в библиотеку. Возникает необходимость дать пользователям возможность специфицировать свои предпочтения [137, 138, 139], развивая возможность определения собственных терминов в рамках некоторого направления научного знания, уточняя и очерчивая круг своих интересов, позволяя организовывать группы пользователей со сходными интересами для возможности отслеживания всей информации по определенным направлениям. Это позволяет лучше понять *информационные потребности* [140, 141, 142] пользователя и облегчить ему поиск нужной информации средствами самой библиотеки.

Развитие технологий интеграции разнообразных источников данных [106, 107, 108, 109, 110, 158] и извлечения из них знаний ставит новые задачи перед семантическими библиотеками. Интеграция данных из разных источников позволяет шире взглянуть на рассматриваемую предметную область, найти новые взаимосвязи и обогатить знания, представленные в семантической библиотеке. Широкое применение *онтологий* позволяет интегрировать данные библиотек с данными из различных источников, основываясь на их семантике. Эти источники не обязательно сами являются библиотеками.

Актуальность проблемы. 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 [115, 117, 121, 122, 126]. При этом средство организации знаний должно быть достаточно универсальным и не требовать глубоких технических познаний.

Говоря далее о произвольных предметных областях, мы будем иметь ввиду 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.

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 интегрирования 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 [47] (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 расширения 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 **основных подзадач** 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) Предложен 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.

3) Определена возможность реализации семантических систем, способных гибко настраиваться под запросы конкретной предметной области.

4) Выполнено упрощение доступа и восприятия больших и сложно структурированных объемов информации пользователем.

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

Научная новизна диссертационной работы заключается в следующем:

- Предложены семантические модели информационных ресурсов и объектов, а также модели терминологического описания в виде тезауруса научных электронных библиотек, отличающиеся гибким описанием семантики не только контента, но и терминологии научной предметной области на основе единой онтологической модели, что позволяет управлять, интегрировать и выполнять навигацию между ними.
- Разработаны 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.

- Предложены алгоритмы решения задач семантического описания произвольной научной предметной области на основе высокоуровневых понятий и формирования ее онтологии.
- Предложены 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.

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.

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 и 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) Предложенные в работе модели, подходы и алгоритмы применены для создания программного обеспечения научных электронных библиотек для

некоторой ПО. 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.

Цели и задачи исследования определили логику изложения материала и структуру диссертационной работы. Она состоит из введения, шести глав и заключения. Каждая глава завершается выводами. Список литературы содержит 160 наименований. В диссертации 15 рисунков, 25 таблиц, приводится 4 приложения.

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.)

Реализация и внедрение результатов работы. Разработанная в диссертации семантическая библиотека для построения научного пространства данных некоторой научной предметной области внедрена, что подтверждено справкой о внедрении.

1. Анализ основных концепций для построения семантических библиотек

В этом разделе приведен краткий обзор и анализ основных концепций и инструментов, которые легли в основу работы. В конце раздела приведен обзор некоторых информационных систем, реализующих функциональность, характерную для семантических библиотек.

1.1.Парадигма Semantic Web

С появлением парадигмы Semantic Web для формализации знаний в различных предметных областях стали активно использоваться онтологии [17, 18, 93, 94, 95]. Как правило, эти знания представлены терминами и атрибутами, отражающими онтологические связи между ними. При формализации знаний появляется возможность проводить семантическую обработку информации. Используются также правила вывода, позволяющие делать заключения об имеющихся или вновь поступивших данных. Данные и онтология с правилами вывода вместе представляют собой базу знаний (knowledge base) некоторой предметной области. Это фактически краткое описание классического подхода к определению базы знаний для любой предметной области. При этом данные предметной области имеют определенную структуру, зафиксированную в онтологии.

Для облегчения процесса автоматической обработки информации в сети некоторое время назад была предложена концепция LOD [45, 46, 47, 127] для размещения и описания данных, опирающаяся на уже имеющиеся наработки парадигмы Semantic Web. Единицей описываемых данных в Semantic Web является ресурс. Каждый ресурс обозначает какой–либо реальный объект, понятие или явление и имеет идентификатор URI (Unified Resource Identificator) [39], который используется для описания знаний о сущности. Эти знания представляются в соответствии с моделью данных RDF (Resource Definition Framework) в виде троек «субъект - предикат – объект». Организация

специального пространства связанных данных Linked Data основывается на практических решениях для публикации и связывания структурированных данных. Термин LOD описывает ту часть данных Linked Data, которая находится в открытом доступе и соответствует основным принципам LOD. Идея LOD оказалась очень привлекательной для различных организаций, многие из которых включили свои источники данных в это облако. Оказались провязаны самые различные типы ресурсов, которые представляют интерес для пользователей библиотек с точки зрения обогащения данных как структурно, так и семантически [62, 118, 119, 127, 155, 156, 157].

Основные принципы LOD призывают использовать для идентификации реальных ресурсов и их цифровых описаний, а также их взаимосвязей, уникальные идентификаторы URI. При этом URI должны обеспечивать доступ к описаниям объектов по протоколу HTTP и представлять описание в виде RDF, которое обеспечивает возможность автоматической обработки информации, а также содержать в своем описании ссылки в виде URI на другие взаимосвязанные ресурсы и их описания. Следуя этим принципам, обеспечивается стандартизированный механизм доступа к данным, и поддерживается глобальный обмен данными независимо от модели данных конкретного узла, обеспечивая возможность их интеграции, основываясь на URI.

Парадигма Semantic Web позволяет структурировать описания ресурсов и представлять их в виде RDF, основываясь на онтологиях. Онтология любой предметной области определяет ее понятия, их тип, структуру, совокупность словарей и классификаторов, которые представляют тезаурус предметной области, обеспечивает доступ к знаниям предметной области в разных источниках. Онтологии позволяют выработать и зафиксировать общее понимание области знания и представить знания в виде, удобном для их автоматизированной обработки, обеспечить возможность получения и накопления новых знаний, а также возможность многократного их использования. Тезаурус же обеспечивает терминологическую поддержку предметной области, облегчает навигацию по разделам предметной области [73, 74, 81, 82, 87, 114] .

Все это послужило предпосылками для лавинообразного роста источников данных, интегрированных в LOD. Традиционные библиотеки, являясь, по сути, центрами данных, с одной стороны не могли игнорировать этот процесс, с другой получили импульс к развитию и обогащению не только своих данных, но и стали трансформироваться и превращаться в хранилища ресурсов самых разнообразных типов, поддерживая их взаимосвязи с другими источниками из LOD. Основная проблема, которая возникает в связи с этим – это гетерогенность онтологий разных источников, которая может препятствовать связыванию данных. Однако, существует множество исследований, которые с разной степенью успеха преодолевают эту проблему.

Таким образом, в свете сказанного выше, библиотеки рассматриваются как хранилища структурированных разнообразных данных с возможностью их интеграции в облако LOD и возможностью определения их тематической направленности. При этом подразумевается, что поддерживается вся традиционная для электронных библиотек функциональность: создание, редактирование, поиск, идентификация ресурсов. Ниже в работе приводится представление онтологии модели данных такой библиотеки, архитектура приложения и описана реализация прототипа этого приложения на примерах разных предметных областей. Исходя из возможностей, предоставляемых применением семантических технологий, пользователь библиотеки получает расширенную функциональность для работы с ресурсами библиотеки, имея возможность описывать область своих интересов в терминах предметной области. Это позволит ему организовывать и описывать собственные коллекции и ресурсы, при необходимости детализируя как описания ресурсов, так и свою область интересов, посредством уточнения ее терминов, создавая свой *микротезаурус*. Где микротезаурус представляет набор выделенных пользователем терминов, возможно дополненных новыми понятиями и связями, как между самими понятиями, так и между этими понятиями и контентом библиотеки.

1.2. Онтологии

Классическое определение онтологии принадлежит Т. Груберу «*Онтология* – это точная спецификация концептуализации» [94]. Под концептуализацией подразумевается абстрактное представление предметной области, тогда как спецификация определяет набор понятий этой предметной области и отношений между ними.

Итак, онтология обозначает совокупность понятий, используемых для описания на концептуальном уровне некоторой предметной области. Это описание представляется в таком виде, чтобы оно могло использоваться для машинной обработки. Состав базовых конструкций онтологии независимо от того, для какой области она создается, одинаков: понятия, атрибуты, отношения, экземпляры. Языком описания онтологий является язык OWL, являющийся стандартом W3C, в свою очередь являющийся расширением схем RDF и RDFS. При конструировании новой онтологии можно импортировать извне уже имеющиеся и использовать их понятия при описании предметной области в рамках создаваемой онтологии.

Онтологии можно классифицировать по различным параметрам (в зависимости от того, с какой целью их классифицируют). Рассмотрим наиболее общую классификацию и определим решаемые нами задачи соответственно этой классификации.

Высокоуровневые онтологии. Такие онтологии описывают наиболее общие понятия (пространство, время, материя, объект, событие, действие и т. д.), которые независимы от конкретной проблемы или области [63].

Ориентированные на предметную область. Во многих дисциплинах сейчас разрабатываются стандартные онтологии, которые могут использоваться экспертами по предметным областям для совместного использования в своей области.

Ориентированные на задачу. Это онтологии, используемые конкретной прикладной программой и содержащие термины, которые используются при разработке программного обеспечения, выполняющего конкретную задачу.

Прикладные онтологии описывают понятия, которые зависят как от онтологии задач, так и от онтологии предметной области. Примером может служить онтология производства автомобилей определенной марки.

В области цифровых библиотек изначально распространение получили так называемые библиографические онтологии, которые описывают типы ресурсов библиотеки, их состав и взаимосвязи. Рассмотрим краткое описание некоторых онтологий, которые часто встречаются в контексте цифровых библиотек, и пример онтологии, ориентированной на построение информационных систем.

Онтология *AKT Reference Ontology* [144] или кратко AKT разработана в целях унификации доступа к библиографической информации в 2003г. И хотя проект был закрыт, данные на сегодняшний момент в AKT предоставлены более чем в 200 источниках таких как: *DBLP* [145], *Citeseer* [146], *CORDIS* [147], *EPSRC* [148], *ACM* [149], *IEEE* [150] и т.д. Объединяет несколько онтологий**,** из которых интерес представляет основная онтология *Portal Ontology*, которая содержит понятия для описания персон и публикаций. Данные разнородны и опираются на очень узкие подмножества этой онтологии. Многие поля, имеющиеся в этой богатой онтологии, остаются незаполненными при описании реальных данных.

Изначально *Dublin Core* (DC) [54] представляет собой набор понятий, используемых для описания разнообразных типов ресурсов, из которых 15 являются обязательными для описания. Элементы DC часто повторно используются, дополняются и конкретизируются в других онтологиях. DC охватывает огромное количество источников, включая DBpedia [11], являющийся фактически центром облака LOD.

Онтология ***FOAF*** [59, 60] уже является практически стандартом для описания персон и их отношений с другими ресурсами. Используется в разнообразных контекстах и может использоваться для описания в любых сценариях с участием персон. Часто также включается и конкретизируется в других онтологиях.

Онтология ***BIBO*** [58] предназначена для описания библиографических данных, включает в себя понятия из других онтологий (таких, как DC и FOAF),

расширяя и конкретизируя их понятия, которые используются при описании ее классов. Содержит 38 видов документов, содержит понятия, необходимые для описания *персон* и *публикаций.* Охватывает такие источники, как Британская Национальная Библиотека, DBpedia и т.д.

Онтология ***Dbpedia***, разработанная в рамках проекта DBpedia, содержит большое количество классов для описания самых разнообразных объектов, включая, например, понятия *публикация* и *персона*. Она также включает в себя понятия из других онтологий, которые используются при описании ее классов. DBpedia является центральным узлом LOD и связывает информацию из самых разных источников, которые ссылаются на нее.

1.2.1.Онтология информационной системы

Согласно онтологическому подходу любая информационная система на концептуальном уровне обладает общим набором понятий, которые описывают понятия любой информационной системы [1, 63, 64, 116]. Наиболее полная онтология для описания информационных систем была представлена в работе [1]. Онтология BWW (Bunge, Wand and Weber ontology) [103] фокусируется на *модели представления*, которая определяет набор понятий, их связей и характеристик, достаточных для описания структуры и поведения информационных систем. Основные аксиомы этой онтологии:

- *Реальный мир* состоит из *Понятий;*
- *Понятия* имеют *Свойства*;
- *Свойства* представляются *Атрибутами*;
- *Понятия* могут быть сложными или простыми;
- *Свойства* могут быть простыми или составными;
- *Свойства* могут быть собственными или приобретенными (наследуемыми);
- каждое *Понятие* имеет *Состояние*;
- *Состояние* определяется *Значением атрибутов* в конкретный момент времени;

- каждое *Понятие* имеет достижимое множество *Состояний*;
- каждое *Понятие* имеет допустимое множество *Состояний*;
- *Понятия* меняются;
- изменение характеризуется изменением *Состояния*;
- *Понятия* находятся в стабильном или не стабильном состоянии;
- *События* влекут изменение *Состояния*:
  - *События* бывают внутренними и внешними;
  - *События* бывают простыми и сложными;
  - *События* могут быть четко определены или слабо определены;
- *Понятия* имеют множество допустимых событий;
- *Понятия* взаимодействуют друг с другом;
  - *Понятия* лежат в основе любой системы;
- *Понятия* имеют независимые свойства;
- *Понятия* взаимодействуют путем изменения значений своих *Атрибутов*;
- *Реальный мир* – это система, в которой каждое *Понятие* взаимодействует хотя бы с одним другим *Понятием*.

Разработчик должен так построить модель на базе этих понятий, чтобы информационная система адекватно отражала ту часть *реального мира*, для автоматизации которой она предназначена. Моделирование онтологии включает следующие шаги:

1. Анализ понятий и объектов реального мира, очерчиваются рамки предметной области;
2. Выделение классов предметной области как группирующих объекты сущностей;
3. Выделение общих для классов свойств;
4. Выделение собственных для классов свойств, анализируя их возможные экземпляры объектов;
5. Выделение специфических свойств классов;

6. Определение набора событий (действий) которые должны выполняться в системе;

7. Выделение подсистем и их понятий для моделируемой области.

1.3. Тезаурусы

В данной работе тезаурусы рассматриваются как множество терминов, которые описывают определенные *понятия предметной области* (используется как синоним, также равнозначное обозначение *концепт предметной области*) и набор различных семантических связей между ними. Наличие таких связей явно указывает на смысловую связь между понятиями предметной области. При этом для представления *понятия предметной области* используются связанные с ним термины.

Главным образом, тезаурус предназначен для поддержки тематической организации информации и поиска. В других подходах для тематической организации и поиска информации могут использовать с одной стороны классификаторы, с другой - инструменты полнотекстового поиска. Подход, основанный на тезаурусах, находится посередине этих подходов и аккумулирует в себе достоинства обоих. Использование классификаторов предполагает наличие предварительно заданного набора тем с поставленным им в соответствие некоторым кодом. Каждый классифицируемый объект должен быть помечен тематическим кодом. Преимуществом классификаторов является то, что по четко обозначенной теме можно получить полный набор объектов, соответствующих тематике классификатора, но для этого требуется предварительная работа по разнесению объектов по соответствующим темам. Полнотекстовый поиск дает возможность задать запросы на естественном языке. Главный недостаток полнотекстового поиска - это то, что для поиска по одной теме, возможно, придется составлять несколько запросов, так как формулировка искомого может быть выражена несколькими способами. Тезаурус же позволяет, используя разные термины, относящиеся к одному концепту/понятию, осуществлять поиск по концептам на естественном языке с помощью этих терминов.

Выделяют основную связь между понятиями тезауруса – это связь между более широкими и более узкими понятиями. Существуют два основных подвида такой связи:

- связь *«часть – целое»*, когда одно понятие, является частью другого понятия;
- связь *«класс - экземпляр»*, когда одно понятие обозначает элемент класса, обозначаемого другим понятием.

Выделяют также связь *«синонимии»*, когда одно понятие может быть выражено несколькими терминами, являющихся синонимами. Среди терминов, относящихся к одному понятию, выделяют предпочтительный термин, который наиболее точно обозначает конкретное понятие. Остальные термины являются менее предпочтительными и являются синонимами. Существуют также горизонтальные (или ассоциативные) связи между понятиями тезауруса, которые выражают дополнительную смысловую связь между ними. Структура понятия тезауруса может состоять из различных элементов и, например, содержать пояснения к ним, раскрывая его смысл и определяя его использование.

При использовании тезауруса для классификации информационных ресурсов, поисковые запросы по коллекциям этих ресурсов могут уточняться на основе терминов используемого тезауруса и его семантических связей, что, несомненно, повышает точность поиска и полноту в смысле пертинентности.

1.3.1.Описание стандарта ISO 2788-1986

Одним из основных документов, регламентирующих формат представления тезауруса, является стандарт ISO 2788-1986 [91, 102]. Он предназначен для описания моноязычных тезаурусов и определяет тезаурус, как набор терминов, связанных между собою соответствующими связями.

Основные атрибуты понятий определенные в этом стандарте приведены ниже:

- *Scope Note (SN)* – комментарий к понятию.
- *Top Term (TT)* – помечает понятия на самом верхнем уровне иерархии.

Выделяются следующие связи между понятиями и терминами:

- *USE* – связь понятия с наиболее предпочтительным термином.
- *Used For (UF)* –.связь понятия с терминами – синонимами.
- *Broader Term (BT)* – иерархическая связь между более общим и более узким понятием.
- *Broader Term Generic (BTG)* – иерархическая связь между более широким понятием и более узким, когда более узкое понятие определяет разновидность более широкого понятия.
- *Broader Term Partitive (BTP)* – иерархическая связь между более широким понятием и более узким, когда более узкое понятие определяет часть более широкого понятия.
- *Related Term (RT)* –горизонтальная или ассоциативная связь между двумя понятиями связанных между собою по смыслу, но не являющихся синонимами.

Выделяют также связи *Narrower Term (NT), Narrower Term Generic (NTG,) Narrower Term Partitive (NTP),* которые являются обратными к связям *BT, BTG и BTP* соответственно.

1.3.2. Описание стандарта ISO 25964

Более современной версией предыдущего стандарта является стандарт ISO 25964 [74, 88, 89, 90, 91, 92]. Модель, описываемая этим стандартом, поддерживает мультиязычные тезаурусы и другие типы словарей. Стандарт содержит рекомендации по установлению и поддержанию взаимного соответствия между несколькими тезаурусами или между тезаурусами и словарями других типов, используемых при информационном поиске. Были представлены рекомендации и схемы модели данных для взаимодействия по сети. В основу схем данных взаимодействия положены модель данных протокола Z39.50 [151, 152] (схема данных Zthes) и рекомендации SKOS, базирующейся на RDF-модели представления понятий.

В новом стандарте типы связей предыдущего стандарта ISO 2788-1986 были существенно урезаны для поддержки интероперабельности и четко определено, что связи устанавливаются на уровне концептов/понятий, а не на уровне их терминов. Используются следующие связи:

- NT – более узкое понятие, взаимообратно BT;
- BT – более широкое понятие, взаимообратно NT;
- USE – связь с понятием, которое используется вместо этого, взаимообратно UF;
- UF – связь с понятием, вместо которого используется этот, взаимообратная связь USE;
- RT – симметричная связь, определяет связанное понятие;
- LE – симметричная связь, определяет связь между лингвистически эквивалентными понятиями.

Определяются следующие типы терминов:

- TT – понятие верхнего уровня;
- NT – понятие неверхнего уровня;
- ND – неосновное понятие;
- NL – фиктивное понятие.

1.4. Концептуальная модель электронных библиотек DELOS

Концептуальная модель электронных библиотек с определениями важнейших представлений об архитектуре, ресурсах и функциональности электронных библиотек была определена в программном документе DELOS [2] (Digital Library Reference Model, DLRM). 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. За подробностями можно обратиться к работе [2].

На основе этой модели предполагается создание конкретных ЭБ, обычно использующих более узкую модель при реализации. Важной особенностью этого стандарта является то, что он не описывает ни логическую, ни, тем более, физическую модель ЭБ, поэтому результат применения стандарта может сильно различаться [153].

Большое внимание в стандарте DELOS DLRM уделяется отделению содержимого ЭБ от определения ЭБ и от СУЭБ. Собственно содержимое ЭБ описывается отдельными стандартами, отображающими наборы метаданных и способы их представления при передаче и хранении. Эти стандарты разрабатываются независимо от DELOS DLRM и решают задачи описания содержимого ЭБ. Использование того или иного стандарта метаданных определяется отдельно в каждой конкретной ЭБ в зависимости от предметной ориентации. В качестве стандартов могут использоваться, например, универсальный Dublin Core, CIDOC-CRM [55], предназначенный для описания музейных объектов, PRISM [56, 57], описывающий публикации.

Область контента представляет собой все объекты, информация о которых доступна в библиотеке и используется для удовлетворения информационных потребностей пользователей. Информационные объекты являются непосредственными составляющими контента библиотеки, в свою очередь являющимися ресурсами электронной библиотеки. Понятие ресурса, в свою очередь, включает в себя их общие характеристики и определяет возможные типы связей для всех типов ресурсов, содержащихся в системе, экземплярами которых являются информационные объекты. Ресурсы и их экземпляры могут объединяться на разных уровнях представления в коллекции. Для однозначной интерпретации ресурса, его описание может быть представлено в виде онтологии, которая может быть довольно сложной, в которой ресурсы могут расщепляться на более мелкие и определяются связи между ними.

1.5. Некоторые реализации семантических библиотек

Разными исследовательскими группами реализованы решения для семантических библиотек.

1.5.1. Greenstone3

Рассмотрим в качестве первого примера популярную систему Greenstone3 [83, 84, 85, 86], которая свободно распространяется, и главной целью которой является создание и поддержка пользовательских коллекций цифровых ресурсов самых разнообразных типов в общественных учреждениях. Система разрабатывается в университете Новой Зеландии под руководством Ян Виттен (Ian H. Witten). 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.

Рассмотрим далее основную функциональность Greenstone. На базе этого программного обеспечения, существует возможность определять структуру описания электронных документов на основе разнообразных метаданных и объединять их в коллекции. Электронные документы могут иметь различные форматы и содержать большие объемы текста и изображений. Поддерживаются такие функции как полнотекстовый поиск, поиск по метаданным, ставшие уже классическими для цифровых библиотек. Доступ к контенту библиотеки может предоставляться как через Интернет, так и на отдельных носителях (компакт - дисках). Система поддерживает многоязычный пользовательский интерфейс и поддерживает обработку коллекций электронных документов для нескольких языков.

Текстовые электронные документы могут быть структурированы согласно содержанию документа. Это удобно при создании индексов для поиска. Поисковые запросы при этом могут состоять, как из одного слова, так и нескольких фраз.

Как и при формировании описания структуры электронных документов, структура коллекции в системе определяется динамически при ее создании. При

этом создается конфигурационный файл, в котором описаны правила использования коллекции. Новые документы в коллекцию включаются, если удовлетворяют условиям, заданным при описании коллекции. Для каждой коллекции создается свой отдельный поисковый индекс, при этом результаты поиска по нескольким коллекциям могут объединяться перед представлением пользователю.

Программное обеспечение состоит из четырех основных модулей

- серверный (Greenstone Server) – его запуск делает компьютер узлом локальной сети Greenstone;
- локальный интерфейс библиотекаря (Librarian Interface) – в этом модуле идет непосредственный ввод книг в электронную библиотеку;
- редактор метаданных (Metadata Set Editor) – здесь возможно редактирование существующих либо создание новых наборов метаданных, создание новых единиц классификации;
- удаленный интерфейс библиотекаря (Remote Librarian Interface).

Фактически интерфейсы поиска и просмотра документов в коллекциях формируются автоматически на основе конфигурации коллекции. Отдельно задаются метаданные, представляемые в интерфейсе просмотра документа, и отдельно помечаются метаданные, которые участвуют в поиске документа в коллекции. Для построения интерфейсов используются также различные классификаторы, которые группируют метаданные по различным признакам.

Из всех изученных систем, Greenstone3 является наиболее близкой по духу к разрабатываемой нами системе. Несмотря на то, что система Greenstone3 достаточно легко устанавливается и настраивается, в ней нет возможности расширения структуры ресурсов коллекций в процессе жизнедеятельности системы. Пользователи системы могут создавать свои коллекции, но возможности специфицировать свою область знаний путем использования тезауруса предметной области и его доопределения путем добавления своих терминов они не имеют. Возможности описания пользователем своей области интересов и ее уточнение, поддержка тематического расширяемого пользователями тезауруса,

возможность динамического расширения описания ресурсов и связей между ними – основной недостаток Greenstone3.

1.5.2. JeromeDL

Семантическая библиотека JeromeDL [3] 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 [4] 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 [5], онтология FOAF [6], модель знаний SKOS [7, 49] для описания таксономий.

Основным недостатком, на наш взгляд, является ориентированность только на библиографические данные, слабая поддержка интеграции данных с другими источниками в рамках системы (в частности, с ресурсами из LOD). При необходимости добавления нового типа ресурсов приходится вносить изменения в систему на программном уровне. Одним из преимуществ этой системы является поддержка, помимо сервиса традиционного атрибутного поиска, сервисов семантического поиска данных на естественном языке, доступ к данным на языке запросов SPARQL [8] для возможности машинной обработки. Следует отметить, что система распространяется бесплатно, но на момент написания этого текста ссылка была недоступна.

1.5.3. Europeana

Одной из глобальных реализованных цифровых библиотек является проект Europeana [9], 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) [10], 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.

Масштаб этой библиотеки одновременно является и одним из препятствий для возможности «индивидуальной тематической» работы пользователя и скорее позволяет причислить ее к глобальным семантическим библиотекам, среди которых также можно указать DBpedia [11], являющуюся ядром облака LOD.

1.6. Постановка задачи и логическая схема исследования

Как видно из приведенного краткого обзора, в последнее время одним из основных направлений работ по созданию информационных систем стало создание их на основе высокоуровневых онтологий. Опираясь на основные концепции DELOS DLRM, адаптировав онтологию BWW [103] для своих целей и используя методику построения модели информационной системы, была поставлена задача *разработки семантической библиотеки* для некоторой научной предметной области.

Такой подход позволяет провести грань между функциями информационной системы, которые обеспечивают взаимодействие с пользователями, поддерживая политику доступа к данным, обеспечивают механизмы интероперабельности системы и процессов обработки данных, и при этом понятие контента библиотеки отделено от описания предметной области, ее ресурсов на семантическом уровне. Документ DELOS описывает наиболее полно на концептуальном уровне все аспекты электронных библиотек, включая их информационные ресурсы. Мы не претендуем на полное соответствие этой

модели, но идеи, изложенные в ней, легли в основу наших исследований. Разрабатывая модель ресурсов информационной системы для разрабатываемой семантической библиотеки, мы хотели получить гибкую систему интеграции различных типов ресурсов с возможностью интеграции с внешними системами. Основные идеи при определении стиля моделирования ресурсов библиотеки при разработке системы были позаимствованы из концепции адаптивных моделей данных, разработанной еще в 90-х годах.

Исходя из вышесказанного, была 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 интегрирования 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.

Для реализации этой системы выделены следующие задачи, реализуемые в данной работе:

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 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 расширения, 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 целей работы 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.

Опираясь на модель понятий, предлагаемую в этой работе, а также идеи Semantic Web и LOD, была разработана персональная открытая семантическая цифровая библиотека LibMeta с системой поддержки работы пользователей с цифровыми ресурсами библиотек и их коллекциями для некоторой предметной области, ограниченной терминологически с помощью тезауруса. Средствами этой библиотеки решаются задачи интеграции данных из различных источников как

включенных в LOD, так и обладающих потенциалом для такого включения посредством самой библиотеки. Разработка велась с учетом концептуальных идей создания и описания библиотек, представленных в стандарте DELOS с учетом накопленного опыта при создании информационных систем.

2. Семантическая библиотека и научная информация

В этом разделе рассмотрим основные концепции, на которых базируется определение семантической библиотеки для некоторой научной предметной области и, основываясь на них, приведем определение разрабатываемой библиотеки и ее контента.

2.1. Этапы развития библиотек

Перед тем как давать определение семантической библиотеки, рассмотрим последовательные этапы эволюции библиотек: электронные, цифровые и семантические библиотеки [2, 3, 10, 11, 12, 50, 51, 52, 53, 67, 68, 69, 70, 77, 96, 105, 133] . Ниже приведем свое определение библиотеки на каждом этапе и выделим в них определение контента.

2.1.1.Электронная библиотека

Формально ***электронная библиотека*** представляет собой тройку объектов $<F, R, A>$, где F –множество функций хранения и поиска, обеспечиваемых информационной системой для обработки объектов множества R. Структура объектов из R представлена фиксированным набором атрибутов $(a_1, \dots, a_k)$, $a_i \in A$. Этот набор будем называть *описанием множества* R или *метаданными множества* R.

Будем называть множество R контентом библиотеки. *Информационным объектом* будем называть любой объект $r \in R$. Тогда описание отдельного информационного объекта будем обозначать как $r(a_1, \dots, a_k)$. При этом значениями a_i могут быть только символьные наборы из некоторого алфавита L. Набор атрибутов и символьные значения этих атрибутов для объекта будем называть метаданными этого объекта. Значения атрибута a_i будем обозначать $r(a_i) \in L^*$, где L^* обозначает множество всех строк (включая пустую строку),

составленных из символов, входящих в L. Множество F состоит из функций вида $f : (a_1, \dots, a_j) \rightarrow In,\ In \subset R, j$ принимает значения от 1 до k.

Фактически контент электронных библиотек представляет собой множество библиографических записей объектов реальной классической библиотеки. В электронных библиотеках не идет речи о цифровом представлении копий реальных объектов, а лишь об их описаниях. В таких описаниях, например, встретив упоминание персоны в одном месте, невозможно точно установить соответствие с ее упоминанием в другом месте.

2.1.2.Цифровые библиотеки

Цифровые библиотеки решают те же задачи поиска и хранения контента, что и *электронные библиотеки*, но существенно расширяют свою функциональность и определение своего контента. Во-первых, контент библиотеки становится мультимедийным. Это значит, что значениями атрибутов ее информационных объектов теперь могут выступать различные мультимедийные объекты, которые доступны для просмотра средствами самой цифровой библиотеки. В качестве мультимедийных объектов могут выступать совокупность аудио, видео, фото и текстовых материалов. Во-вторых расширяется функциональность за счет решения некоторых задач интеграции как метаданных, так и медийных объектов из внешних источников, доступных по сети.

При этом формальное определение представляет собой уже набор объектов $<F, R, A, Mul>$, где F, R, A определяются, так же, как и в определении электронных библиотек. Mul представляет собой множество доступных мультимедийных объектов и $r(a_i) \in (Mul\ U\ L^*)$.

Множество F дополняется функциями вида $g : (X, a_1, \dots, a_j) \rightarrow Out$, где $Out \subset R(X)$ и $Out \subset R$, где X – внешний источник, а множество объектов $R(X)$ может быть описано набором атрибутов $(a_1, \dots, a_k)$. Функции g предназначены для решения вопросов интеграции данных из внешних библиотек.

2.1.3.Семантические цифровые библиотеки

Семантические цифровые библиотеки 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 интегрировать данные различных типов. Обеспечивается интероперабельность с другими системами, не обязательно являющимися библиотеками, так как основной задачей семантических технологий является предоставление метаданных в машиночитаемом формате.

Формально семантическая цифровая библиотека – это $<F, R, A, \Phi>$, где F, A определяются так же как и в определении *цифровых библиотек*. Контент библиотеки $R = R_1 U R_2 U \ldots U R_s$ представляет собой множество типов информационных ресурсов системы, для каждого их которых определен свой набор атрибутов $(a_{i1}, \ldots, a_{ik})$. Такое определение не означает исключение мультимедийных объектов, а подчеркивает обыденность мультимедийных объектов в семантических библиотеках. То есть $M \in R$, и должно пониматься как добавление нового типа контента *«мультимедийный объект»* в библиотеку со своим набором атрибутов и отношений, каждый объект которого является абстрактным представлением реального объекта из множества M. Значения атрибутов $c(a_{ij}) \in (L^* U R)$. L^*, как и прежде содержит область значений строковых атрибутов из A. Φ задает множество условий, накладываемых на представление контента, которое может, например, содержать ограничения, накладываемые на форматы значений $r(a_{ij})$.

Важную роль в определении семантических библиотек при описании их контента играют онтологии. Онтология модели контента фактически задается $<R, A, I>$, где множество R выступает как множество понятий онтологии, множество атрибутов A также содержит подмножество отношений между понятиями, а I – задает множество функций интерпретации, заданных на понятиях и отношениях.

Таким образом, множества *R, A, I* задает описание структуры контента библиотеки.

2.2. Основные свойства семантических библиотек

Основным свойством семантических библиотек является возможность структурирования их разнообразного контента и возможность связывания данных из разных источников, что в свою очередь, несомненно, отражается на качестве данных контента.

Выделим основные свойства семантической библиотеки, которые, на наш взгляд, являются определяющими для рассматриваемой системы:

- семантическая библиотека представляет собой интеграционный узел для разных источников данных, которые обогащают и пополняют ее набор данных;
- контент библиотеки описывается на семантическом уровне, что позволяет достичь лучшего взаимодействия между источниками данных;
- контент библиотеки может иметь разную степень гранулярности структуры в зависимости от рассматриваемых начальных условий при построении библиотеки;
- семантическое описание контента и его уровень гранулярности не зависят от технических характеристик реализации информационной системы библиотеки и могут определяться вне зависимости от конкретной реализации;
- понятийное описание контента библиотеки поддерживается его тезаурусом, который ограничивает предметную область ресурсов библиотеки терминологически.

2.3. Информационные системы в контексте семантических библиотек. Общая терминология

Выделив выше модель контента семантической библиотеки и ее основные характеристики, мы отделили определяющее понятие контента семантической библиотеки от понятия реализующей библиотеку информационной системы [12, 63, 66, 116]. Такой подход позволяет наращивать функциональность системы, добавлять новые подсистемы или изменять уже имеющиеся при неизменных остальных частях.

Информационная система *IS* задается набором подсистем *F* для решения задач обработки ее контента *C*, *IS* = (*F, C)*. Тогда информационная система представляется как организация совокупности своих подсистем $F = U F_i$ и своего контента *C*. Каждая из этих подсистем описывается своей предметной онтологией и тогда можно представить ***онтологию информационной системы*** *OnIS* интеграцией онтологий ее подсистем и онтологии ее контента $OnIS = OnF \; U \; OnC$, где $OnF = U \; OnF_i$ - объединение онтологий подсистем, $OnC = <R, A, \Phi>$ – онтология контента. При описании онтологий информационных систем и ее модулей обычно опираются на абстрактные онтологии высокого уровня, определяя ее ключевые сущности [1, 13].

Каждый вид деятельности, поддерживаемый в рамках информационной системы, обеспечивается ее отдельной подсистемой. При разработке каждой подсистемы выделяются задачи, решаемые данной подсистемой, определяются набор и структуры данных, определяется набор пользовательских интерфейсов и граф навигации по ним. При этом набор и структура данных подсистемы определяются без привязки к конкретной предметной области. При интеграции с онтологией контента онтология подсистемы уточняется за счет встраивания понятий контента как подклассов ее понятий данных. Такой подход к интеграции онтологий носит уточняющий характер и направлен с одной стороны на расширение онтологии информационной подсистемы, а с другой стороны ограничивает ее применение рамками заданной предметной области.

Приведем основные виды задач, реализующиеся в информационной системе, предназначенной для конструирования семантической библиотеки:

- описание контента информационной системы;
- реализация задач интеграции данных из внешних источников;
- поддержка коллекций;
- поиск и навигация по объектам системы;
- поддержка пользователей;

Разбиение на подсистемы не является единственно возможным. Границы подсистем не могут быть строго определены. В системе существует область общих определяющих понятий, которые рассматриваются как принадлежащие нескольким подсистемам, в зависимости от того, какие процессы выполняются в конкретной подсистеме. Так или иначе, каждая из этих систем взаимодействует с понятиями, определяющими контент этой библиотеки. Например, в перечисленных подсистемах можно рассматривать как единую подсистему реализации задач интеграции данных из внешних и внутренних источников. С другой стороны, из подсистемы качества можно выделить отдельно систему выявления дубликатов. Это деление диктуется конкретной реализацией.

2.4. Научная информация как контент семантической библиотеки

Построение обобщенной модели научной предметной области представляет 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 [48, 66, 67, 71, 72, 101, 114, 134, 135]. Это позволит построить общую модель научных знаний, в рамках которой могут интегрироваться различные источники данных. Такой подход позволяет структурировать и связать различные *ресурсы*, извлечь из них и контекстуализировать разнообразные *данные*, превращая их в *знания*.

С использованием обобщенной модели возможна реализация модели контента семантических библиотек, способной гибко настраиваться под запросы предметной области. Одной из целей обобщенного подхода является упрощение доступа и восприятия больших и сложно структурированных объемов информации пользователем [136]. Этот подход не является оптимальным для всех задач, решаемых в рамках некоторой научной предметной области, но, по крайней мере, позволяет структурировать имеющиеся знания на формальном уровне для дальнейшего использования.

Критерии научности информации строго не определены и на этот счет существуют различные точки зрения. Научная информация, по ГОСТ 7.0-99 [14],: логически организованная информация, получаемая в процессе научного познания и отображающая явления и законы природы, общества и мышления. Опираясь на это определение, можно выделить несколько основных свойств, которыми в совокупности обладает научная информация: истинность, интерсубъективность и системность [15, 16].

Мы не будем претендовать на то, что знания о предметной области, описанные в соответствии с нашими предложениями, являются всеобъемлющими. Оценивать - задача экспертов, наша же задача - предоставить удобный инструмент для анализа имеющейся информации.

Критерий интерсубъективности говорит о том, что у всех исследователей, изучавших одну и ту же предметную область в одних и тех же условиях, получится один и тот же результат. Мы хотим предложить такую модель организации информации о предметной области, которая является общезначимой для всех исследователей.

Системность научной информации подразумевает опору на исследование разнообразных зависимостей. Спецификой такой информации является четкая структура организации научных данных в иерархические структуры, пронизанные горизонтальными связями. Как следствие, соответствие этим основным критериям обеспечивает достаточно однозначную интерпретацию научного знания различными исследователями.

Основная проблема представления научной информации состоит в сложности используемых понятий и отношений между ними, и, что самое главное, они подвержены более частому изменению структур данных, что неизбежно приводит к необходимости внесения существенных доработок в уже имеющиеся решения.

Говоря о научной информации, имеет смысл разделять понятия научных данных и научных ресурсов в рамках научных предметных областей.

2.4.1.Научные ресурсы

В то же время, особенностью электронных и других источников научных данных является то, что, несмотря на сильное различие в назначении, они предоставляют похожие ресурсы для любой предметной области науки. То есть, информационные ресурсы в разных предметных областях представляются часто одними и теми же объектами: научные публикации, ученые, работающие в этой отрасли, научные организации, проекты, гранты, опыты, образцы, экспериментальные установки и другие. При этом непосредственно научные данные могут извлекаться из этих ресурсов.

2.4.2.Научные данные

Одновременно с построением обобщенной модели возникает необходимость ограничения ее в рамках конкретной предметной области науки. Для этого вводится набор понятий, используемых для описания этой предметной области. Соответствующие термины предметной области связывают с этими понятиями. Чаще всего эти термины организованы в виде некоторой таксономии с поддержкой связей между ними. Структура этой таксономии может варьироваться по сложности в зависимости от моделируемой области и представлять собой при необходимости полноценный тезаурус со всем богатством связей. В дальнейшем будем говорить о тезаурусах как о средстве организации понятий (знаний). Представленные в таком виде термины могут употребляться для обработки имеющихся ресурсов. При этом между понятиями и

ресурсами возникают связи. Под научными данными предметной области будем понимать совокупность понятий научной предметной области и выявленных связей между ними и ресурсами.

Отдельно стоит упомянуть о том, что тезаурус предметной области может быть как результатом работы экспертной группы, так и построен автоматизированными средствами. Вопросы составления тезаурусов предметной области выходят за рамки этой работы, как и обсуждение методов выявления связей.

2.4.3.Научные знания

Совокупность научных ресурсов, их экземпляров, терминов тезауруса, всех явных и неявных связей между ними образуют общую картину научных знаний предметной области.

Самый простой набор правил выявления неявных связей определен в самом стандарте онтологического представления информации [17]. 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.

Онтологическое описание фактически является формальной основой для представления научных знаний и учета свойств интерсубъективности [18].

2.5. Адаптивный подход к описанию контента

Формирование модели с перечисленными свойствами фактически подводит нас к построению онтологии контента семантической библиотеки, близкой, по сути, к высокоуровневым онтологиям [19] для предметных областей науки. Такие онтологии описывают наиболее общие понятия, независимые от конкретной проблемы или области.

В этом смысле как нельзя лучше подходит *адаптивная модель данных* [20, 21] для описания научных ресурсов, которая позволяет не ограничиваться при

разработке строго очерченным набором ресурсов. Такая модель данных подходит для определенного круга задач, решение которых реализуется на базе построения довольно сложных частных моделей. Применение адаптивной модели позволяет понизить сложность (размерность) как самой модели данных, так и разрабатываемых на ее основе систем. Получаемые модели более абстрактны, состоят из меньшего количества понятий с более простыми связями и не привязаны к определенным предметным областям. Использование этой модели данных делает возможной динамическую трансформацию и интерпретацию модели данных в приложении, позволяет настраивать решения под определенную предметную область. Фактически появляется возможность воспроизвести и поддерживать в процессе развития описание различных структур и процессов, используемых в рассматриваемой предметной области. Такой подход позволяет значительно улучшить качество обработки и поиска поступающих ресурсов и данных в рамках ограниченной предметной области не только за счет использования ее тезауруса, но и за счет гибкости описания представления имеющихся ресурсов.

2.5.1.Разработка общей терминологии

Очевидно, что при разработке общей идеологии нужно иметь в виду, что на глобальном уровне концептуализации должны присутствовать понятия, используя которые можно описать структуру знания любой научной предметной области. Выше мы разделили научные данные и научные ресурсы. В свою очередь для полноты необходимо ввести связи между данными и ресурсами. Перечислим основные понятия, необходимые для описания научных знаний:

- *Ресурс*. Определяет источник научной информации и задает описание информационных ресурсов. Из-за разнообразия возможных ресурсов на глобальном уровне нельзя использовать фиксированный перечень ресурсов, так как их набор и степень детализации могут меняться от условий задачи.
- *Объект*. Представляет экземпляр ресурса конкретное описание объекта.

- *Понятие*. Определяет единицу научной информации и описывает научные данные. Множество таких понятий составляет набор сведений о предметной области.
- *Связь между понятиями*. Описывает связь между двумя понятиями.
- *Тезаурус*. Множество терминов и связей между ними.
- *Связь между понятиями и объектами*. Описывает взаимосвязь между Понятием и Объектом и наоборот. При этом понятие не обязательно должно быть связано с объектом. Но каждый объект являющийся экземпляром научного ресурса, относящегося к предметной области, должен быть связан с каким либо понятием. Совокупность этих связей и представляет научные знания.
- *Правило*. Определяется на основе связей между объектами и понятиями. Позволяет вывести знания, не определенные явно с помощью понятий, на глобальном уровне концептуализации.

Совокупность этих понятий задает обобщенное представление модели научной предметной области и вместе с данными составляют пространство научного знания рассматриваемой области.

С помощью этих понятий возможен переход на следующий уровень концептуализации контента библиотеки – предметный. На предметном уровне для некоторой области происходит описание существенных для этой области понятий для научных ресурсов, данных и правил вывода новых знаний.

2.6. Выводы

Опираясь на основные концепции, изложенные выше, выделим набор основных задач, которые должна решать ***персональная открытая семантическая цифровая библиотека***:

1. библиотека должна поддерживать возможность использования медийных объектов или ссылки на них при описании своих объектов, включая текст, аудио, видео файлы или любую их комбинацию. Это требование отражается в названии словом «цифровая»;

2. типы используемых ресурсов и связи между ними должны быть описаны средствами системы в рамках определенных понятий, составляющих семантическое описание ресурсов контента библиотеки. Набор этих понятий тематически ограничивается терминами предметной области из некоторого тезауруса. Это требование отражаться в названии словом «семантическая»;
3. библиотека должна являться интеграционным узлом, предоставляя возможность связывания своих данных с данными из разных источников, которые включены в облако источников данных Linked Open Data (LOD). Также должна обеспечиваться возможность извлекать данные этой библиотеки в машиночитаемом формате. Это требование отражается в названии словом «открытая»;
4. пользователи библиотеки должны иметь возможность организовывать свои коллекции по интересующему их научному направлению, добавляя новые термины в предметный тезаурус, уточняя, таким образом, область своих интересов. Так же пользователи должны иметь возможность осуществлять поиск не только среди объектов в рамках системы, но и по источникам данных без необходимости использования специальных знаний для поисковых запросов. Это требование отражается в названии словом «персональная».

Основные требования, предъявляемые при этом к контенту системы, - *универсальность*, *структурированность*, *адаптируемость* не противоречат этим свойствам и обеспечивают поддержку настраиваемого хранилища метаданных для объектов и расширяемый набор информационных ресурсов. *Универсальность* обеспечивает описания ее типов ресурсов и объектов независимо от предметной области и области интересов пользователей. *Структурированность* описания обеспечивает поддержку связей между различными типами ресурсов как внутри системы, так и вне ее, исходя из определений LOD. *Адаптируемость* описания ресурсов обеспечивает возможность добавления новых свойств и связей в

процессе развития системы и обеспечивает настройку пользовательских интерфейсов под эти изменения.

Далее рассмотрим основные понятия информационной модели, которые позволяют выстроить систему для конструирования семантической библиотеки таким образом, чтобы выполнялись перечисленные выше требования.

3. Информационная модель семантической библиотеки

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, дадим 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 [104, 106, 107, 108, 109, 110, 111, 112, 113, 123]. На основе этих определений описываются основные процессы такие, как, например, интегрирование данных из разных источников, категоризация/классификация, отображение разных моделей данных источников на заданную предметную область, построение классов эквивалентности и т.д.

3.1. Основные свойства модели

Перечислим основные свойства разрабатываемой модели семантической библиотеки научного пространства знаний [22, 23, 24]

- *направленность* – модель отображает некоторую обобщенную систему представления знаний научной предметной области, которая имеет целью возможность моделирования научных знаний любой направленности;

- *конечность* – модель отображает оригинальную предметную область в ограниченном количестве ее состояний, так как возможности моделирования ограничены;
- *упрощенность* – модель отображает только необходимые стороны предметной области для конкретной решаемой задачи;
- *адекватность* - модель должна адекватно отображать объективные закономерности представления знаний научной предметной области;
- *наглядность* – модель представляется набором основных понятий, отношений и свойств;
- *доступность и технологичность* – модель должна быть доступной для исследования и для машинной обработки;
- *информативность* – модель должна содержать достаточную информацию о пространстве знаний научной предметной области и должна давать возможность получить новую информацию из имеющейся;
- *полнота* - в модели должны быть учтены все основные понятия, связи и отношения, необходимые для обеспечения цели моделирования не зависимо от конкретной предметной области;
- *целостность* - модель реализует некоторую систему представления знаний предметной области;
- *замкнутость* - модель реализует замкнутую систему необходимых основных понятий, связей и отношений, достаточных для представления пространства научных знаний предметной области;
- *адаптивность* - модель может быть приспособлена к различным предметным областям;
- *эволюционируемость* – модель должна иметь возможность изменения или расширения представления пространства научных знаний предметной области.

Несмотря на большой список свойств модели, становится понятным, что построение идеальной модели научного пространства знаний отражающей все возможные его аспекты, практически сложно достижимая задача. Но при учете

перечисленных свойств при ее построении, мы можем 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. Эти свойства можно рассматривать как качественные характеристики разрабатываемой модели.

3.2. . Модель контента библиотеки

3.2.1.Основные определения

Введем следующие понятия и обозначения для описания модели контента:

- *o – информационный объект, OI = {o_1, ..., o_n}* – множество информационных объектов,
- *r – информационный ресурс, R = {r_1, ..., r_m}* – множество информационных ресурсов (типов информационных объектов),
- *TYPE(o)= r* – функция соответствия объекту информационного ресурса;
- *a – атрибут, A = {a_1, ..., a_k}* множество атрибутов информационных ресурсов, *Z(a)* – множество значений атрибута *a*.
- *IsPLURAL(a)* – функция значение которой выбирается из множества *{true, false}, IsPLURAL(a) = false,* если атрибут *a* может иметь единственное значение, т.е. *Z(a)* состоит из одного элемента, и *IsPLURAL(a) = true* в противоположном случае. Далее будем говорить, что функция *IsPLURAL(a)* определяет тип значений атрибута.
- A_D – множество атрибутов, значением которых может являться любое значение из некоторого множества *D*, в котором не может быть информационных ресурсов системы, т.е. $\forall a_i,\ a_i \in A_D : Z(a_i) = d,\ d \in D,\ D \notin OI$.
- A_I – множество атрибутов, значением которых может являться любой объект из *OI*

  $\forall a_i, a_i\ A_I : Z(a_i) = o,\ o \in OI,$

$A_D \vee A_I = A,\ A_D \wedge A_I = \emptyset$

- v – *вид атрибута,* $V = \{v_1, \ldots, v_t\}$ – множество видов атрибутов. При этом атрибут a может быть нескольких видов, т.е $\forall a, a \in A : VIEW(a) = \{v_k\}, v_k \in V, 1 \leq k \leq t$ далее для простоты будем обозначать вид атрибута $VIEW(a) = v_k$.
- Для атрибутов определены область значений $DOM(a) = R', R' \in R$ и область определения $RAN(a) = X : X \in D$ или $X \in OI$.
- SA – это множество таких подмножеств атрибутов, в каждое из которых входят хотя бы по одному атрибуту каждого вида, каждое такое подмножество называется набором атрибутов

  $\forall v, A' v \in V, A' \in SA\ \exists a, a \in A'\ VIEW(a)=v, v \in V.$

  При этом одному информационному ресурсу могут соответствовать несколько наборов атрибутов, т.е. $SET(r) = A'', A'' = \{A'\}, A' \in SA$, где SET функция соответствия набора атрибутов ресурсу. Из $SET(r) = A'$, где $A'= \{a_1, \ldots, a_t\}$, следует что $\forall o \in OI$, такого что $TYPE(o)= r, o = (Z(a_1), \ldots, Z(a_t))$. Или можно сказать, что если информационному ресурсу соответствует некоторый набор атрибутов, то все объекты являющиеся экземплярами этого ресурса, описываются как набор значений этих атрибутов.

По своим функциям атрибуты можно делить на следующие пересекающиеся *виды атрибутов*: *идентифицирующие*, *озаглавливающие*, *обязательные*, *классифицирующие*, *поисковые*, *описательные* [71]. В процессе обработки данных каждая задача в системе использует определенный тип атрибутов. Например, идентифицирующими атрибутами ресурса будем называть подмножество атрибутов набора атрибутов, необходимых и достаточных для однозначного определения каждого из его информационных объектов. 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. Для описания формальной общей картины эти тонкости пока

можно опустить, полагая, что мы можем включить в набор атрибутов столько однотипных атрибутов, сколько требуется.

Представление информационного ресурса *P(A', v)* определяется множеством используемых в нем атрибутов, соответствующих ресурсу (используется, например, для автоматической генерации описания объекта для пользователя или какого-то агента). Фактически представляет выборку значений атрибутов объекта некоторого ресурса для заданного вида атрибутов.

Определение 1. *Контент библиотеки* $C = <R, A, OI>$ определяется типами ее информационных ресурсов *R*, описанных связанными с ними наборами атрибутов *A* и набором входных данных, определяющих информационные объекты *OI*, которые являются непосредственно объектами, хранящимися в библиотеке.

3.2.2.Основные утверждения

1. Атрибут a_i может быть только одного типа: $\forall a_i \in A,\ A_D \wedge A_I = \emptyset\ ,\ => \ a_i \in A_D \vee a_i \in A_I$.
2. Множеству атрибутов информационного ресурса соответствует хотя бы один набор атрибутов:

   $\forall r = \{a_1, \ldots, a_k\},\ r \in R,\ \exists A' \subset SA: SET(r) = A',\ a_j \in A',\ 1 \le j \le k.$
3. Любому информационному объекту соответствует определенный набор атрибутов:

   $\forall o_i \in OI\ \ \exists A' = SET(r)\ ,\ A' \subset S :\ r = TYPE(o_i)\ .$
4. Если атрибуты информационного ресурса входят в разные наборы атрибутов, то этому ресурсу соответствует несколько наборов атрибутов :

   $\forall r,\ r \in R,: a_i \in A_1',\ a_j \in A_2',\ a_j,\ a_i \notin A_1' \wedge A_2',\ a_j \neq a_i,\ \exists A_1' = SET(r),\ A_2' = SET(r).$
5. Тип атрибута позволяет определить область значений этого атрибута:

   $\forall a_i \in A_D\ ,\ \exists D_D = DOM(a_i),\ D_D \subset D\ ,\ Z(a_i) = d_D,\ d_D \in D_D,$

$\forall a_j \in A_I$, $\exists D_I = DOM(a_j)$, $D_I \subset OI$, $Z(a_j) = o_i$, $o_i \in D_I$.

6. Область определения атрибутов задается ресурсами в чьи наборы атрибутов они входят:

   $\forall a_i \in A'$, $\exists r$: $SET(r) = A' <=> RAN(a_i) = r$,

   $SET(r_1) = A'$, $SET(r_2) = A''$ $<=> RAN(a_i) = r_1 \vee r_2$.

7. Если область значений атрибута состоит из объектов, то соответствующий им информационный ресурс также является ресурсом системы:

   $\forall a_j \in A_I$, $\exists r = TYPE(o_i)$, $Z(a_j) = o_i$.

8. Если некоторый ресурс является подклассом информационного ресурса, то он сам является ресурсом системы:

   $\forall r_2$, $IsRe(r_1, r_2)$, $r_1 \in R => r_2 \in R$.

9. Если представление описывает некоторый ресурс, то это также представление для любого объекта соответствующего информационного ресурса:

   $\forall o_i \in OI$ $\exists P(A', v) = P(SET(r), v) =$ $P(SET(TYPE(o_i)), v)$.

10. Если некоторому ресурсу соответствует набор атрибутов, то этот набор также соответствует любому его подресурсу:

    $\forall r_1, r_2 \in I_R$, $IsRe(r_1, r_2)$, $\exists A' \subset S$ $A' = SET(r_1) = SET(r_2)$.

11. Если представление описывает некоторый ресурс, то это также представление для любого его подресурса:

    $\forall r_1, r_2 \in R$, $IsRe(r_1, r_2)$, $\exists P(A', v_i) = P(SET(r_1), v_i) = P(SET(r_2), v_i)$.

12. Если представление описывает некоторый ресурс, то это также представление для любого объекта его подресурса:

    $\forall o_i \in OI$, $TYPE(o_i) = r_2$, $IsRe(r_1, r_2)$, $\exists P(A', v_i) = P(SET(r_1), v_i) =$

$= P(SET(r_2), v_i) = P(SET(TYPE(o_i)), v_i)$.

13. Если представление определено для некоторого набора атрибутов, то оно определено для любого информационного ресурса соответствующего этому набору:

    $\forall A'$, $A' = SET(r_1) = SET(r_2) =>$ $P(A', v_i) = P(SET(r_1), v_i) = P(SET(r_2), v_i)$.

3.2.3.Примеры построения запросов

1) Извлечь все объекты определенного информационного ресурса из системы:

   $\exists\, o_i \;:\; o_i \in OI \wedge TYPE(o_i) = r \wedge r \in R$.

2) Извлечь все объекты определенного информационного ресурса, у которых задано значение o_i атрибута a_j , и o_i является информационным объектом:

   $\exists\, o_i : o_i \in OI \wedge TYPE(o_i) = r \wedge r \in R \wedge a_i \in A_I \wedge Z(a_j) = o_i \wedge o_i \in D_I \wedge D_I = DOM(a_j) \wedge D_I \subset OI.$

3) Извлечь все объекты определенного информационного ресурса, у которых задано определенное значение o_i атрибута a_j, и при этом o_i не является информационным объектом:

   $\exists\, o_i : o_i \in OI \wedge TYPE(o_i) = r \wedge r \in R \wedge a_i \in A_D \wedge Z(a_j) = o_i \wedge o_i \in D_D \wedge D_D = DOM(a_j) \wedge \neg(D_D \subset OI)$.

3.3. Модель тезауруса предметной области

Построение информационной модели тезауруса, соответствующей стандарту ISO 25964, позволит использовать ее как каркас для построения форматизированного представления терминологической модели знаний представляемых тезаурусом для любой предметной области [72, 73, 74, 75, 76].

3.3.1.Основные определения

Введем основные понятия и обозначения для описания тезауруса:

- C_p *– множество концептов,*
- $T = P_T \cup N_T$ - *множество* вербальных *терминов* для концептов и их синонимы, где P_T множество главных терминов, а N_T их синонимы
- $C_R = H \cup V$ - *множество связей* между концептами, где H множество горизонтальных и V - множество вертикальных связей;
- $CONCEPT(t_i) = c$ где $CONCEPT$ это функция, ставящая любому термину $\forall t_i \in T$ в соответствие его концепт $c \in C_p$.

- *TERM(c) = { t_1, … ,t_k}* где *TERM* это функция, которая сопоставляет концепту множество терминов
- *PTERM(c) = t_j* где *PTERM* - это функция, которая сопоставляет концепту один предпочитаемый термин
- *NTERM(c) = { t_1, … , t_k}* где *NTERM* - это функция, которая сопоставляет концепту синонимичные термины.
- *RELOBJ(c) = {o_1, … , o_k}* где *RELOBJ* - это функция, которая сопоставляет концепту информационные объекты.
- *RELCONC(o) = {c_1, … , c_k}* где *RELCONC* - это функция, которая сопоставляет объекту концепты тезауруса.
- Также по аналогии с моделью контента библиотеки вводятся понятия
  - A_{TH} – атрибут тезауруса,
  - SA_{TH} – набор атрибутов тезауруса,
  - *SET(TH)* – функция, которая сопоставляет тезаурусу набор атрибутов
  - функция *IsPLURAL(a)*, $a \in A_{TH}$ определяет тип атрибута тезауруса;
  - функция *VIEW(a)*, $a \in A_{TH}$ определяет вид атрибута
  - функция *DOM(a)* , $a \in A_{TH}$ определяет область значений;
  - функция *RAN (a)* , $a \in A_{TH}$ определяет область определения.

Эти понятия используются для создания расширенного описания структуры концепта тезауруса.

Определение 2. *Тезаурус библиотеки* $TH = < C_p\ T,\ C_R,\ SA_{TH}>$ определяется концептами C_p и их терминами T, связями C_R между ними. Набор концептов C_p и их терминов T, составляющих терминологическое описание предметной области, строго задан. А также определяется набором дополнительных атрибутов тезауруса SA_{TH}.

3.3.2.Основные утверждения

1) Если $TERM(c) = \{ t_1, \dots, t_k \}$, то $\exists\ j: PTERM(c) = t_j$ и $NTERM(c) = \{ t_1, \dots, t_{j-1}, t_{j+1}, \dots, t_k \}$, т.е. только один термин концепта является предпочитаемым, остальные его синонимы.
2) Иерархические связи транзитивны и асимметричны

   $v \in V$ иерархическая связь между терминами, тогда

   - для любых терминов c_1, c_2, c_3 выполняется $v(c_1, c_2) \wedge v(c_2, c_3) \Rightarrow v(c_2, c_3)$
   - для любых терминов c_1, c_2 из $v(c_1, c_2) \not\Rightarrow v(c_2, c_1)$
3) Горизонтальные связи ассоциативны

   $h \in H$ горизонтальная связь между терминами, тогда

   - для любых терминов c_1, c_2, c_3 выполняется $h(c_1, c_2\ c_3) = h(c_1, h(c_2, c_3)) = h(h(c_1, c_2), c_3)$
4) Атрибут теауруса a_i может быть только одного определенного типа: $\forall a_i \in A_{TH}$, $A_D \wedge A_I = \emptyset$, $\Rightarrow\ a_i \in A_D \vee a_i \in A_I$.
5) Структура концепта может варьироваться (дополняться)

   Набор атрибутов тезауруса SA_{TH} дополняет структуру концепта этого тезауруса, т.е. если $SA_{TH} = \{a_1, \dots, a_i\}$, то описание концепта дополняется значениями атрибутов из этого набора $c_i = \{Z(a_1), \dots, Z(a_i)\}$

Также выполняются утверждения касающиеся атрибутов тезауруса, аналогичные утверждениями для атрибутов информационных ресурсов.

3.3.3.Примеры запросов

1) Извлечь связанные с концептом тезауруса объекты

   $\exists o_i : o_k \in OI \wedge RELOBJ(c) = \{o_1, \dots, o_k\} \wedge r \in R$.
2) Извлечь связанные с концептом тезауруса объекты определенного типа

   $\exists o_i : o_k \in OI \wedge TYPE(o_i) = r \wedge r \in R \wedge RELOBJ(c) = \{o_1, \dots, o_k\}$.
3) Извлечь объекты, связанные с концептами, которые в свою очередь связаны между собой горизонтальными связями

$\exists\, o_i : o_i \in OI \wedge RELOBJ(c_1) = \{o_{11}, \ldots, o_{1n}\} \wedge RELOBJ(c_j) = \{o_{j1}, \ldots, o_{jm}\} \wedge h(c_1, c_j), h \in H$.

4) Извлечь объекты, связанные с концептами, которые в свою очередь связаны между собой иерархическими связями

   $\exists\, o_i : o_i \in OI \wedge RELOBJ(c_1) = \{o_{11}, \ldots, o_{1n}\} \wedge RELOBJ(c_2) = \{o_{j1}, \ldots, o_{jm}\} \wedge v(c_1, c_j), v \in V$.

5) Извлечь множество объектов связанных с данным объектом посредством концептов тезауруса

   $\exists\, o_i : o_i \in OI \wedge RELCONC(o) = \{c_1, \ldots, c_k\} \wedge RELOBJ(c_j) = \{o_{j1}, \ldots, o_{jn}\}$.

6) Извлечь множество терминов концептов связанных с объектом

   $\exists\, t_i : t_i \in T \wedge RELCONC(o) = \{c_1, \ldots, c_k\} \wedge TERM(c_j) = \{ t_{j1}, \ldots, t_{jk}\}$.

7) Извлечь множество терминов из объектов связанных с данным посредством концептов тезауруса

   $\exists\, t_i : t_i \in T \wedge RELCONC(o) = \{c_1, \ldots, c_k\} \wedge RELOBJ(c_j) = \{o_{j1}, \ldots, o_{jn}\} \wedge \wedge RELCONC(o_j) = \{c_{j1}, \ldots, c_{jk}\} \wedge TERM(c_l) = \{ t_{l1}, \ldots, t_{lk}\}$.

3.4. Модель интеграции

Исходя из основных понятий, 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. Введем функцию *simp(a)*, которая ставит в соответствие каждому атрибуту тип его значения и возвращает *0* для простых типов, и *1* для объектных типов.

Формально подсистема интеграции I_T представляется тройкой $I_T = <G, \{S_i\}, \{M_i\}>$, 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, $1 \le i \le n$, где n количество источников данных.

Схему ресурса как источника данных S, так и контента библиотеки G можно представить в виде графа, который включает объекты и отношения. Каждый объект может быть связан отношением с другим объектом, значения которого представлены простыми типами данных (строки, числа, даты) или отношениями с другими объектами, значения которых соответствуют некоторым ресурсам. При этом для отображения ресурса мы можем использовать его представление Z_s, то есть выбрать не полный набор его атрибутов и отношений с другими объектами для отображения на схему G, но который при этом обязательно включает в себя набор атрибутов, значения которых позволяют однозначно идентифицировать объект в системе.

Семантические связи между представлением Z_s некоторого ресурса, соответствующего S, и ресурсом, соответствующего G, определяют элементы отображения источников данных на модель контента библиотеки. Семантическая связь m_i, определяющая отображение элемента источника данных s на элемент контента библиотеки $g \in G$, обозначим как $(m_i(Z_s) => s \leftrightarrow g)$.

Если $s \leftrightarrow g$, где $g = r$, то для построения рабочего отображения ресурса r на источник S достаточно построить множество отображений

$$M = \{(m_r(Z_s) => s \leftrightarrow g = r) \vee \{(m_i(Z_s) => s \leftrightarrow g = a_i)\} \mid r(a_1,.., a_n) \in R, a_i \in I(r)\},$$

В случае, когда $g = a$, соответствие $s \leftrightarrow^* g$ будем называть *точным,* если связываемые элементы совпадают по смыслу, $s \leftrightarrow^+ g$ – *избыточным,* если s шире по смыслу g, $s \leftrightarrow^\wedge g$ – *неточным,* если s уже по смыслу g.

Результатом работы отображения ресурсов на схему источника данных является множество связей между информационными объектами библиотеки и данными подключенного источника, т.е.

$$M : (S, R) \rightarrow \{sameAs(o_S, o_R) \vee seeAlsoAs(o_S, o_R)\}$$

3.4.1.Набор стандартных определений операций для построения отображения

Операция eq. Если *simp(a)=0, g = a* и $s \leftrightarrow^{*} g$, то будем использовать операцию *eq(s, g)* для оценки совпадения значений соответствующих атрибутов объекта из источника со значением атрибута информационного объекта из LibMeta.

Операция ext. В случае, когда *simp(a)=0, g = a* и $s \leftrightarrow^{+} g$ будем использовать операцию *ext(s, g)*, чтобы извлекать значение атрибута объекта из источника для поиска в нем соответствующего вхождения значения атрибута информационного объекта из LibMeta.

Операция inc. В случае, когда *simp(a)=0, g = a* и $s \leftrightarrow^{\wedge} g$ будем использовать операцию *inc(s, g)*, чтобы извлекать значения соответствующих атрибутов объекта из LibMeta для поиска в нем соответствующего вхождения значения атрибута объекта из источника.

Операция split. Если *simp(a)=0,*$g_i = a_i$, $g_j = a_j$ и для различных g_i и g_j выполняется $s_k \leftrightarrow^{+} g_i$, $s_k \leftrightarrow^{+} g_j$, то данные из источника должны рассекаться, чтобы соответствовать данным конечной схемы контента, для этого определяется пара операций, *{split(*s_k*,* g_i*), split(*s_k*,* g_i*)}*.

Операция app. Если *simp(a)=0,* $g_k = a$, для различных s_i и s_j и выполняется $s_i \leftrightarrow^{+} g_k$, $s_j \leftrightarrow^{+} g_k$, то данные из источника должны объединяться, чтобы соответствовать данным конечной схемы контента, для этого определяется пара операций, *{ app(*s_k*,* g_i*), app(*s_k*,* g_i*)}*.

Операция norm. При необходимости, при *simp(a)=0, g = a*, в операциях *eq, ext, inc, split, app* могут использоваться вспомогательные операции для нормализации данных согласно описаниям соответствующих атрибутов в системе, например, *s'=norm(s), g'=norm(g), inc(s', g').*

Набор операций *op = {eq, ext, inc, split, app, norm}* выполняют задачи преобразования данных из терминов исходной системы в термины источников и.

Операция see. В случае, когда *g = r* и *s* ↔ *g,* будем определять связь ресурса из источника с информационным ресурсом объекта из LibMeta. Мы не претендуем на построение полного отображения ресурсов или их иерархии, как в других

подходах при интеграции схем [25, 26, 106, 107, 110, 111, 112, 113, 120]. Нам достаточно связи, которая означает, что экземпляры ресурсов в рамках библиотеки и источника данных могут относиться к одному и тому же объекту [124, 125]. Такой подход позволяет нам строить частичные отображения ресурсов, достаточные для однозначной идентификации объектов и определения связи между ними.

Операция res_eq. В случае *simp(a)=1*, $g = a$ и $s \leftrightarrow^{*} g$, будем использовать операцию ***res_eq*** *(s, g)* для оценки совпадения объектов, значений соответствующих атрибутов.

Операция add_att. Если для некоторого *s* не существует g , такого что $g = a$ и $s \leftrightarrow g$, будем использовать операцию *add_att(s, g)* для добавления такого $g=a$ в схему ресурса, для которого *simp(a)=0* или *simp(a)=1*.

Операция add_res. Если для выполнения операции *add_att* не существует g, такого что $g = r$ и $s \leftrightarrow g$, *для simp(a)=1*, будем использовать операцию *add_res(s, g)* для добавления нового ресурса в схему такого что $g = r$ и $s \leftrightarrow g$.

Операции *add = {add_att, add_res}* делают подсистему интеграции настраиваемой под любой источник данных с возможностью обогащения уже имеющихся в системе данных. Динамическое доопределение модели *G* фактически включает в себя этап анализа ресурсов интегрируемого источника данных и расширение или уточнение исходной модели *G* путем расширения множества *R* или множества *A*. Возможность выполнения этих операций обеспечивается благодаря принятой адаптивной модели данных системы.

3.4.2.Построение отображения

Процесс построения отображения можно разделить на несколько основных этапов:

- Подключение источника данных. Каждый источник данных характеризуется соответствующим уникальным *URL* адресом и некоторым набором параметров, необходимых для доступа к данным.

- Проводится предварительный анализ доступной из источника информации, в частности определяются типы его ресурсов и их свойства, участвующие в интеграции. Результатом первого этапа становится определение той части схемы источника S_i, по которой будут извлекаться данные. При необходимости выполняются операции *add* для настройки модели данных контента.
- Определение типов ресурсов библиотеки, соответствующих типам ресурсов источников. Для каждого ресурса источника, определенного его схемой, извлеченной на этом этапе, ставится в соответствие ресурс библиотеки. Результатом этого этапа становится установление связи между ресурсом библиотеки и ресурсом источника с помощью операции ***see***, которая декларирует, что существуют экземпляры этих ресурсов, соответствующие одному и тому же объекту реального мира. На базе определенных связей *{see(r, r_s)}* на следующем этапе проходит отображение атрибутов.
- Для каждого ресурса определяется отображение атрибутов на соответствующие им свойства ресурса источника данных. В первую очередь строится отображение для идентифицирующих атрибутов, являющихся обязательными, затем для остальных. Для каждой пары *(r, r_s)*,такой, что определенна операция *see(r, r_s)*, определяется тип связи $s \leftrightarrow g$, где $g = a_i$, и определяется набор операций *op(s, g)*.

Благодаря такому построению отображения мы получаем набор правил, по которым мы можем представить каждый найденный объект в источнике в рамках понятий нашей 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.

3.4.3.Запросы к интегрируемым источникам данных

После построения отображения становится возможным осуществление поисковых запросов по источникам данных. То есть, любой экземпляр ресурса источника данных, для которого построено отображение, может являться ответом на запрос пользователя.

Определим $I_T=(G', \{S_i\}, \{M_i\})$ как систему интеграции данных, $L=\{L|S_i \in I_T\}$ как интерпретацию источников в I_T,. Если в источнике данных S_i есть множество объектов $\{obj\}$, которые могут являться экземплярами ресурса $r \in G'$, тогда по запросу пользователя q в терминах G' можно построить по интерпретации L запрос q_L, такой что ответом на него будут именно объекты множества $\{obj\}$, представленные в терминах G'.

Отдельной проблемой представления консолидированного представления ответа пользователю из различных источников данных является выявление дублирующих объектов. Для этой цели используются значения идентифицирующих атрибутов. Конечно, не существует оптимального набора таких атрибутов для любого типа ресурса, но, по крайней мере, эксперт может подобрать оптимальный набор для ресурсов в рамках заданной предметной области, на основе которых и принимается решение о том, являются ли объекты из разных источников идентичными или нет.

Описанная модель интеграции, может, не является оптимальной для любого случая и имеет свои недостатки, характерные для общих подходов описанных выше. Но в рамках очерчиваемых постановкой задачи границ и внешних источников данных удалось достигнуть приемлемых результатов в процедурах по упрощению подключения поиска и навигации по таким источникам.

3.5. Дополнительные определения

Введем недостающие определения в информационную модель для выделения дополнительных семантических связей в рамках контента библиотеки

Определение 3. *Семантические метки* $SM = \{sm_i\}$ информационного объекта – это термины, которые не попали в тезаурус, но являются необходимыми для специфицирования тематики информационного объекта. Семантические метки не связаны, в отличие от терминов тезауруса, связями между собой или с терминами тезауруса, но дают возможность дополнительного тематического разделения информационных объектов в рамках предметной области.

Определение 4. *Коллекция информационных объектов* $C = \langle IO, T, SM, S \rangle$ представляет собой набор объектов, объединенных на основе совокупности признаков:

1) по их термину тезауруса предметной области;
2) по семантическим меткам;
3) по источнику данных, из которого поступили объекты.

В коллекцию могут входить объекты различных типов ресурсов, заданных при описании контента библиотеки. При этом коллекции по каждому признаку могут формироваться автоматически, либо пользователем по произвольным признакам.

Определение 5. *Семантически значимыми связями библиотеки* $P = \{P_i\}$ назовем связи, определенные между контентом библиотеки, понятиями ее предметной областью (тезаурусом), семантическими метками и объектами источника данных. Выделим следующие основные связи:

- $P_1(t, o)$ – термин тезауруса – информационный объект;
- $P_2(o, t)$ – информационный объект – термин тезауруса;
- $P_3(r, s)$ – информационный ресурс – класс объектов источника, где информационный ресурс – это общее определение для информационных объектов, хранящихся в системе, т.е. информационные объекты, являются экземплярами информационных ресурсов;
- $P_4(a, s_a)$ – атрибут информационного ресурса – свойство класса источника;
- $P_5(o, o_s)$ – информационный объект – экземпляр класса из источника данных;
- $P_6(sm, o)$ – семантическая метка – информационный объект;
- $P_7(o, sm)$ – информационный объект – семантическая метка.

На основе введенных явных связей можно определить связи, которые назовем *неявными значимыми связями* (то есть заданными по некоторым определенным заранее правилам) между семантическими метками и терминами тезауруса и объектами как самой библиотеки, так и экземплярами связанных данных из источников:

- $P_8(sm, t) \leftarrow P_6(sm, o) \wedge P_2(o, t)$ семантическая метка – информационный объект – термин тезауруса;

- $P_9(t,sm) \leftarrow P_1(t, o) \wedge P_7(o, sm)$ термин тезауруса – информационный объект – семантическая метка;
- $P_{10}(sm, o_s) \leftarrow P_6(sm, o) \wedge P_5(o, o_s)$ семантическая метка – информационный объект – экземпляр класса из источника данных;
- $P_{11}(t, o_s) \leftarrow P_1(t, o) \wedge P_5(o, o_s)$ термин тезауруса – информационный объект – экземпляр класса из источника данных.

Для представления онтологии был выбран язык описания онтологий OWL. Такая онтология состоит из классов, свойств классов и индивидов. В терминах OWL P_1 *инверсивно* P_2, P_6 *инверсивно* P_7, P_8 *инверсивно* P_9, P_{10} *инверсивно* P_{11}.

3.6. Выводы

В описании информационной модели семантической библиотеки были введены понятия для описания содержимого библиотеки для некоторой предметной области. Эти понятия позволяют сконструировать описание любых типов информационных ресурсов для этой области. При этом согласно определению информационные объекты, являющиеся непосредственно содержимым библиотеки, имеют распределенную природу, что означает, что данные могут поступать из различных источников и агрегировать информацию об информационном объекте из различных источников, непосредственно сохраняя данные в самой библиотеке или сохраняя ссылки на идентичные объекты в источниках данных.

Для описания ресурсов, составляющих контент конкретной предметной области, предлагается использовать понятия, общие для любой из них. То есть набор понятий, формирующих описание контента библиотеки, должен быть настолько универсальным, чтобы мог адаптироваться под нужды конкретной области.

Контент библиотеки тесно связан с тезаурусом, который поддерживает родственные связи различных типов, как между концептами, так и между концептами и информационными объектами. Это позволяет реализовать гибкий настраиваемый поиск, результатом которого будет сбалансированный список

объектов по предметной области. На основе одного и того же тезауруса определяются коллекции самых разнообразных типов ресурсов. Такой подход чрезвычайно полезен для создания раздельных пользовательских коллекций.

Так как одной из основных задач, решаемых в рамках библиотеки, как было сказано выше, является интеграция данных из различных источников, такой 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 определенным требованиям в рамках введенных понятий, а именно:

1. источник данных содержит информацию о некоторых объектах;
2. каждый объект в источнике однозначно идентифицируется;
3. информация об объекте задается значениями его атрибутов;
4. каждый объект может быть связан с другим объектом, находящимся как в источнике, так и за его пределами.

Этим требованиям удовлетворяют в частности источники из LOD, которые рассматриваются в качестве основных поставщиков данных в этой работе.

4. Построение онтологии семантической библиотеки научного пространства знаний

Описание онтологии научного пространства знаний 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.

Исходя из вышесказанного, тезаурус пространства знаний – это полный систематизированный набор терминов о какой-либо области знаний и больше относящийся к лексике, используемой в конкретной области, тогда как онтология описывает ресурсы предметной области и их взаимосвязи. Для каждой предметной области набор ресурсов может отличаться как по формату, так и по набору самих ресурсов.

Онтология научного пространства знаний - это сложная многоуровневая система понятий, описывающих ресурсы и объекты предметной области, концептов, терминов и связей между ними, характеризующаяся открытой

иерархической и динамичной структуризацией и служащая как для хранения знаний и их структуризации, так и для добычи новых.

4.1. Построение многоуровневой онтологической модели научной предметной области

Для возможности построения онтологии научного пространства знаний необходимо придерживаться следующей последовательности шагов при конструировании семантической библиотеки в рамках предложенного подхода.

1. На основе введенной модели задается набор информационных ресурсов, используемых в библиотеке. Для этого необходимо представить описания содержимого будущей библиотеки в терминах предложенной модели.
2. Осуществляется окончательная настройка структуры тезауруса. На базе определенных классов согласно определению задаются используемые связи между терминами, расширяется при необходимости описание термина, определяются связи с ресурсами системы.
3. Для выбора семантических меток можно использовать дополнительные словари по предметной области или оставить возможность их доопределения позднее.
4. Наполнение онтологии данными согласно описанию модели пространства знаний заданных на первых трех этапах.

После выполнения последовательности шагов 1 – 3, мы фактически получаем *упрощенную* модель предметной области, описанную 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 предложенной ниже 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.

4.1.1.Базовые понятия сущностей предметной области

В соответствии с определениями, рассмотренными в этой и предыдущей главе, были введены основные классы онтологии. Указываемая при описании определения класса таблица его свойств устроена следующим образом:

- В столбце «Название» указывается название свойства и в скобках его XML-идентификатор (т.н. квалифицированное имя элемента – префикс пространства имен и локальное имя после двоеточия); этот идентификатор используется при описании данных. Он также определяет URI элемента в соответствии с его пространством имен. В столбце «Комментарий» указывается текстовое пояснение смысла свойства и, возможно, формата его значений. Могут указываться также такие характеристики OWL-свойства, как суперсвойство и обратное отношение.
- В столбце «Тип значений» приводится указание требуемого типа значений свойства: примитивного типа (строка, число, дата или другие допустимые в RDF типы данных XML Schema), либо некоторого класса (в таком случае – это ссылка, либо двустороннее отношение), определенного в данной схеме или в одной из схем, от которых она зависит.

- После указания типа значений в том же столбце указывается допустимое количество значений свойства – мощность свойства:
  - [0..*] означает «от 0 до бесконечности» – свойство является множественным и необязательным для указания (факультативным).
  - [1..*] означает «от 1 до бесконечности» – свойство является множественным и обязательным для указания.
  - [0..1] означает «от 0 до 1» – свойство допускает не более одного значения, является необязательным для указания (факультативным).
  - [1..1] означает «ровно 1» – свойство допускает ровно одно значение, является обязательным для указания.

Были введены следующие суперклассы для разделения используемых классов:

- ***ClassWithURI*** – группирует классы, экземпляры которых представляют собой уникально-идентифицируемые объекты.

Название свойства	**Комментарий**	**Тип значений**
Дата создания (lb:dateOfCreation)		Дата и время (xsd:dateTime) [1..1]
Дата последнего обновления (lb:dateOfCreation)		Дата и время (xsd:dateTime) [1..1]
Создатель (dc:creator)		Пользователь (lb:User) [1..1]
URI (lb:uri)	Уникальный идентификатор объекта	Строка (xsd:string) [1..1]

- ***ClassWithoutURI*** – соответственно группирует классы экземпляры которых не представляют собой уникально-идентифицируемые объекты.
- ***SecureClass*** – группирует классы, используемые для определения прав доступа к сущностям описываемой предметной области

Далее, исходя из определения контента, вводятся классы:

- ***InformationResource*** (информационный ресурс библиотеки), который содержит общую информацию о типе ресурса, название, *URI* и информацию об используемом наборе атрибутов для описания структуры ресурса.

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI)

Название свойства	**Комментарий**	**Тип значений**
Название (dc:title)		Строка (xsd:string) [1..1]
Метка (lb:label)		Строка (xsd:string) [1..1]
Описание (dc:description)		Строка (xsd:string) [0..1]
Связанная таксономия (lb:taxonomy)	В качестве значения указывается словарь или классификатор	Таксономия (lb:Taxonomy) [0..*]
Эквивалентный класс (owl:equivalentClass)	Ссылка на URI-идентификатор эквивалентного класса.	Строка (xsd:string) [0..*]
Набор атрибутов (lb:resourceAttributeSet)		Набор атрибутов (lb:AttributeSet) [1..*]

- ***InformationObject*** (информационный объект библиотеки), который фактически представляет собой экземпляр некоторого ресурса и по составу атрибутов соответствует набору атрибутов связанного с ним ресурса. Для описания соответствующих значений для информационного объекта имеется многозначное свойство *value*, значениями которого являются экземпляры вспомогательного класса ***AttributeValue***, содержащие информацию о конкретном значении объекта и соответствующем атрибуте.

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI)

Название свойства	**Комментарий**	**Тип значений**
Описание (dc:description)		Строка (xsd:string) [0..1]
Тип ресурса (lb: informationResourceType)	В качестве значения указывается экземпляр информационного ресурса	Информационный ресурс (lb:InformationResource) [1..1]
Значение (lb: objectAttributeValue)		Значение атрибута (lb:AttributeValue) [0..*]

Элементы классификатора/словаря (lb:taxons)		Код элемента (lb:Taxon) [0..*]
Смотри также(owl:sameAs)	Ссылка на другой аналогичный экземпляр возможно содержащий дополнительную информацию	Строка (xsd:string) [0..*]

- ***ResourceAttribute*** – класс, который описывает атрибут, элемент описания информационного ресурса, который содержит информацию о типе значений, указывает на область применения атрибута в рамках системы. Может иметь значения *поисковый* (участвует в формировании поисковых форм), *идентифицирующий* (является обязательным) и *описательный* (содержит информацию об описываемом объекте в человекочитаемом виде).

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI)

Название свойства	**Комментарий**	**Тип значений**
Название (lb:title)		Строка (xsd:string) [1..1]
Описание (dc:description)		Строка (xsd:string) [0..1]
Метка (lb:label)		Строка (xsd:string) [1..1]
Видимость (lb:visibility)		Да/нет (xsd:boolean) [1..1]
Тип значения (lb: resourceAttribyteType)		Тип атрибута (lb: ResourceAttributeType) [1..1]
Вид атрибута (lb: attributeView)		Вид атрибута (lb:AttributeView) [1..1]
Многозначность (lb:isMultiple)		Да/нет (xsd:boolean) [1..1]
Смотри также (owl:equivalentProperty)	Ссылка на другое аналогичное свойство	Строка (xsd:string) [0..*]

- ***ResourceAttributeSet*** набор атрибутов экземпляров класса ***ResourceAttribute***, группирующий атрибуты, соответствующие одному представлению ресурса.

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI)

Название свойства	**Тип значений**
Название (lb:title)	Строка (xsd:string) [1..1]
Описание (dc:description)	Строка (xsd:string) [0..1]
Метка (lb:label)	Строка (xsd:string) [1..1]
Атрибуты (lb:attributes)	Атрибуты ресурса (lb:ResourceAttribute) [1..*]

- ***Taxonomy*** группирует классы для описания базовой структуры таксономий, например, таких как словари, классификаторы

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI)

Название свойства	**Тип значений**
Название (lb:title)	Строка (xsd:string) [1..1]
Описание (dc:description)	Строка (xsd:string) [0..1]
Метка (lb:label)	Строка (xsd:string) [1..1]

- ***Vocabulary*** класс для контролируемых словарей. В самом общем понимании - это набор терминов некоторой предметной области и правил их использования для описания информации. Наиболее простая, но в то же время часто используемая форма контролируемого словаря – это плоский словарь. Чаще всего плоские словари используются для группировки некоторого набора ключевых терминов и наиболее употребимых фраз и добавления к ним расшифровок, определений, описаний.

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI, lb:Taxonomy)

Название свойства	**Тип значений**
Термины словаря (lb: hasVocabluaryTerm)	Термины словаря (lb: VocabularyTaxon) [1..*]

- ***Classifier*** – это класс для описания классификаторов (рубрикаторов), который представляет собой набор терминов (рубрик) и связей между ними, образующих древовидную структуру. Классификаторы используются для тематической или иной классификации ресурсов с целью упрощения их поиска.

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI, lb:Taxonomy)

Название свойства	**Тип значений**
Термины словаря (lb: hasClassifierTerm)	Термины словаря (lb: ClassifierTaxon) [1..*]

- ***Taxon*** базовый элемент таксономии

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI)

Название свойства	**Тип значений**
Код (lb:code)	Строка (xsd:string) [1..1]
Описание (dc:description)	Строка (xsd:string) [0..1]
Определение (lb: definition)	Строка (xsd:string) [1..1]
Сокращенное название (lb:acronym)	Строка (xsd:string) [0..*]
Приоритет (lb:order)	Целое число (xsd:int) [0..1]

- ***VocabularyTaxon*** – класс соответствующий элементам контролируемого словаря

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI, lb:Taxon)

- ***ClassifierTaxon*** – класс, соответствующий элементам словаря (рубрикатора)

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI, lb:Taxon)

Название свойства	**Тип значений**
Более широкий термин (lb:broader)	Термин классификатора (lb:ClassifierTerm) [0..1]
Более широкий термин (lb:broader)	Термин классификатора (lb:ClassifierTerm) [0..*]
Связанный термин (lb:related)	Термин классификатора (lb:ClassifierTerm) [0..*]

Исходя из определения тезауруса, согласно описанному ранее стандарту, вводятся классы:

- ***Thesaurus*** (тезаурус предметной области), содержит в себе общую информацию о тезаурусе: название и авторов (организации и персоны). Наличие этой сущности позволяет загружать готовые тезаурусы, не смешивая их с теми, что уже, быть может, есть в системе.

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI)

Название свойства	**Тип значений**
Название (lb:title)	Строка (xsd:string) [1..1]
Описание (dc:description)	Строка (xsd:string) [0..1]
Метка (lb:label)	Строка (xsd:string) [1..1]
Множество тематических групп концептов (lb:conceptGroups)	Группа концептов (lb:ConceptGroup) [0..*])
Множество концептов связанных с тезаурусом (lb:concepts)	Концепт (lb:Concept) [0..*])
Множество атрибутов тезауруса (lb:thesaurusAttributeSet)	Набор атрибутов (lb: ThesaurusAttributeSet) [1..1])

- ***Concept*** – сущность, содержащая информацию о понятиях тезауруса.

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI)

Название свойства	**Комментарий**	**Тип значений**
Код (lb:code)		Строка (xsd:string) [1..1]
Комметарий (lb:comment)		Строка (xsd:string) [1..1]
Дополнительный атрибут (lb:conceptAttributeValue)		Значение дополнительного атрибута (lb: ConceptAttributeValue) [0..*]
Предпочитаемый термин (lb:prefferedTerms)	Дескриптор	Термин (lb: Term) [1..1]
Дополнительный термин (lb:nonPrefferedTerms)	Синонимы дескриптора	Термин (lb: Term) [0..*]
Тематическая группа терминов (lb:theme)		(lb: ConceptGroup) [1..1]
Связанный тезаурус (lb:thesaurus)		(lb: Thesaurus) [1..1]

- ***ConceptGroup*** – тематическое разделение понятий тезауруса.

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI)

Название свойства	**Тип значений**

Код (lb:code)	Строка (xsd:string) [1..1]
Комметарий (lb:comment)	Строка (xsd:string) [1..1]
Связанный тезаурус (lb:thesaurus)	(lb: Thesaurus) [1..1]
Заголовок темы (lb:title)	Строка (xsd:string) [1..1]
Множество концептов связанных с тематической группой (lb:concepts)	Концепт (lb:Concept) [0..*])

- ***HierarchicalRelation*** – иерархические связи, определяющие древовидную структуру словаря. Содержит атрибуты, определяющие связи в соответствии со стандартом (BT, BTG, BTP).

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI, lb:ThesaurusRelation)

Название свойства	**Тип значений**
Связанный концепт (lb:firstConcept)	Концепт (lb:Concept) [1..1]
Связанный концепт (lb:secondConcept)	Концепт (lb:Concept) [1..1]
Тип связи (lb:familyRelationType)	Тип связи (lb:FamilyRelationType) [1..1]
Тип взаимно обратной связи (lb:familyRelationReverseType)	Тип связи (lb:FamilyRelationType) [0..1]

- ***FamilyRelation*** – горизонтальные связи. Они задают родственные отношения между понятиями и позволяют находить публикации по похожим тематикам. Содержит также атрибуты, определяющие связи в соответствии со стандартом (NT, NTG, NTP).

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI, lb:ThesaurusRelation)

Название свойства	**Тип значений**
Главный термин (lb:parentConcept)	Концепт (lb:Concept) [1..1]
Подчиненный термин (lb:childConcept)	Концепт (lb:Concept) [0..*]
Тип связи (lb:hierarchicalRelationType)	Тип связи (lb: HierarchicalRelationType) [1..1]
Тип взаимно обратной связи (lb:hierarchicalRelationReverceType)	Тип связи (lb:HierarchicalRelationType) [0..1]

- ***Term*** – термины понятия. Общий класс, объединяющий дескрипторы и синонимы.

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithoutURI)

Название свойства	**Тип значений**
Концепт (lb:concept)	Концепт (lb:Concept) [1..1]
Значение (lb:value)	Строка (xsd:string) [1..1]
Признак дескриптора (lb:isPreffered)	Булевое (xsd:boolean) [1..1]
Видимость (lb:visibility)	Булевое (xsd:boolean) [1..1]

В этот класс объектов добавлен атрибут *visibility* – свойство, отвечающее за видимость термина. Имеет два значения – *global* и *private*, глобальная и приватная область видимости соответственно. Этот атрибут введен для решения проблемы множественных терминологий – разные люди могут называть одни и те же объекты по-разному (пусть даже эти названия будут похожи). Для того чтобы каждому пользователю было комфортно работать в системе, ему дается возможность создавать свои термины, если таковых нет в глобальной части тезауруса. Эти термины он может связывать с другими терминами из глобальной части и размечать ими свои публикации. Таким образом, если два пользователя создали в своих локальных репозиториях удобные для них ключевые слова, разметили ими свои публикации и связали эти ключевые слова с одним и тем же термином из глобального тезауруса, то они смогут находить и получать публикации друг друга, пользуясь при этом своими терминологиями.

- ***ThesaurusAttribute*** – класс, который описывает атрибут, элемент описания концепта тезауруса, который содержит информацию о типе значений, указывает на область применения атрибута в рамках системы.

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI)

Название свойства	**Комментарий**	**Тип значений**
Название (lb:title)		Строка (xsd:string) [1..1]
Описание (dc:description)		Строка (xsd:string) [0..1]
Метка (lb:label)		Строка (xsd:string) [1..1]

Тип значения (lb: thesaurusAttribyteType)		Тип атрибута (lb: ThesaurusAttributeType) [1..1]
Многозначность (lb:isMultiple)		Да/нет (xsd:boolean) [1..1]

- ***ThesaurusAttributeSet*** набор атрибутов экземпляров класса ***ThesaurusAttribute***, группирующий атрибуты, которые расширяют описание концепта тезауруса .

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI)

Название свойства	**Тип значений**
Название (lb:title)	Строка (xsd:string) [1..1]
Описание (dc:description)	Строка (xsd:string) [0..1]
Метка (lb:label)	Строка (xsd:string) [1..1]
Атрибуты (lb:attributes)	Атрибуты тезауруса (lb:ThesaurusAttribute) [1..*]

Исходя из определений семантических меток и коллекций, вводятся классы:

- ***SemanticTag*** – класс семантических меток, который обладает следующими свойствами:

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithoutURI)

Название свойства	**Комментарий**	**Тип значений**
Значение (lb:value)	Краткое название семантической метки	Строка (xsd:string) [1..1]
Описание (lb:description)	Расширенное описание семантической метки	Строка (xsd:string) [1..1]

- ***Collection*** – класс определенных коллекций

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI)

Название свойства	**Комментарий**	**Тип значений**
Название (lb:name)	Название коллекции	Строка (xsd:string) [1..1]
Описание (lb:definition)	Описание коллекции	Строка (xsd:string) [1..1]
Таксономия (lb:base)	Таксономия, на базе которой определяется	Таксономия (lb:Taxonomy) [1..1]

	коллекция	
Элементы коллекции (lb:items)		Информационные объекты (lb:InformationObject) [0..*]
Ресурсы (lb:resources)	Типы ресурсов включаемых в эту коллекцию	Информационные ресурсы (lb:InformationResource) [1..*]

Исходя из определения задачи интеграции, вводятся классы:

- ***DataSource*** – общий класс для подключаемых источников данных из LOD

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithURI)

Название свойства	**Комментарий**	**Тип значений**
Название (lb:name)	Название источника	Строка (xsd:string) [1..1]
Описание (lb:definition)	Описание источника	Строка (xsd:string) [1..1]
URL (lb:url)	точка входа для извлечения данных	Строка (xsd:string) [1..1]
Типы ресурсов (lb:resourceMapping)	Содержит информацию о типах ресурсов, отображаемых на этот источник, и соответствующие классы источника	Отображение ресурсов (lb:ResourceMapping) [1..*]

- ***ResourceMapping*** – класс, содержащий информацию об отображаемых на источник данных информационных ресурсах библиотеки:

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithoutURI)

Название свойства	**Комментарий**	**Тип значений**
Ресурс (lb:resource)	Тип ресурсов, отображаемых на этот источник	Строка (xsd:string) [1..1]
Название класса (lb:dataSourceClass)	Ссылка на соответствующий класс источника данных	Строка (xsd:string) [1..1]
Отображение атрибутов (lb:attributeMappings)	Содержит информацию об отображении соответствующих ресурсу	(lb:AttributeMapping) [1..*]

	атрибутов.	

- ***AttributeMapping*** – класс, содержащий информацию об отображаемых на источник данных атрибутах из набора атрибутов, соответствующего информационному ресурсу библиотеки:

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithoutURI)

Название свойства	**Комментарий**	**Тип значений**
Атрибут (lb:attribute)	Атрибут, отображаемый на этот источник	Атрибут (lb:Attribute) [1..1]
Атрибут источника (lb:dataSourceProperty)	Ссылка на соответствующее свойство класса источника данных	Строка (xsd:string) [1..1]

4.1.2.Детализация понятий сущностей предметной области

Рассмотрим специфические связи между классами онтологии, которые позволяют гибко конструировать наборы понятий второго уровня для конкретной предметной области и ее пространства знаний.

Атрибуты информационных ресурсов определяют структуру ресурсов и их содержание. Для представления различных типов атрибутов в онтологии поддерживаются следующие подклассы класса ***ResourceAttribute,*** которые определяют тип возвращаемого значения:

- ***ResourceAttributeXML*** – класс атрибутов, соответствующий некоторому информационному ресурсу, значение которого представляется в *xml* – синтаксисе.
- ***ResourceAttributeText*** – класс атрибутов, соответствующий некоторому информационному ресурсу, значение которого представляется в виде текста
- ***ResourceAttributeTaxonomy*** – класс атрибутов, соответствующий некоторому информационному ресурсу, значение которого представляется в виде таксона определенного словаря или классификатора

- ***ResourceAttributeString*** – класс атрибутов, соответствующий некоторому информационному ресурсу, значение которого представляется в виде строки.
- ***ResourceAttributeObject*** – класс атрибутов, соответствующий некоторому информационному ресурсу, значение которого представляется в виде информационного объекта определенного типа входящего в контент библиотеки
- ***ResourceAttributeNumber*** – класс атрибутов, соответствующий некоторому информационному ресурсу, значение которого представляется в виде числа
- ***ResourceAttributeDate*** – класс атрибутов, соответствующий некоторому информационному ресурсу, значение которого представляется в виде даты
- ***ResourceAttributeHref*** – класс атрибутов, соответствующий некоторому информационному ресурсу, значение которого представляется в виде ссылки на ресурс в сети
- ***ResourceAttributeFile*** – класс атрибутов, соответствующий некоторому информационному ресурсу, значение которого представляется в виде некоторого файла.

Для расширения описания базовой версии тезауруса, а именно структуры концептов тезауруса также поддерживается иерархия классов для дополнительных атрибутов концептов, которая во многом аналогична описанной иерархии классов атрибутов для информационных ресурсов и включает в себя подклассы суперкласса ***ThesaurusAttribute*** такие как

- ***ThesaurusAttributeText*** – класс атрибутов, расширяющий описание структуры концептов, соответствующих определенному тезаурусу, значения которых представляются в виде некоторого текста.
- ***ThesaurusAttributeTaxonomy*** – класс атрибутов, расширяющий описание структуры концептов, соответствующих определенному тезаурусу, значения которых представляются в виде элемента определенного классификатора или словаря

- ***ThesaurusAttributeString*** – класс атрибутов, расширяющий описание структуры концептов, соответствующих определенному тезаурусу, значения которых представляются в виде строки.
- ***ThesaurusAttributeObject*** – класс атрибутов, расширяющий описание структуры концептов, соответствующих определенному тезаурусу, значения которых представляются в виде некоторого информационного объекта входящего в контент библиотеки.
- ***ThesaurusAttributeNumber*** – класс атрибутов, расширяющий описание структуры концептов, соответствующих определенному тезаурусу, значения которых представляются в виде числа.
- ***ThesaurusAttributeHref*** – класс атрибутов, расширяющий описание структуры концептов, соответствующих определенному тезаурусу, значения которых представляются в виде ссылки.
- ***ThesaurusAttributeFile*** – класс атрибутов, расширяющий описание структуры концептов, соответствующих определенному тезаурусу, значения которых представляются в виде некоторого файла.
- ***ThesaurusAttributeConcept*** – класс атрибутов, расширяющий описание структуры концептов, соответствующих определенному тезаурусу, значения которых представляются в виде других концептов тезауруса (определяют связи между концептами не поддерживаемые явно в системе).

Каждый из этих классов в соответствии с парадигмой наследования поддерживаемой в OWL содержит свойства, описанные для суперкласса ***ThesaurusAttribute***.

4.1.3. Описание экземпляров сущностей предметной области

Если классы из предыдущих разделов использовались для описания онтологической модели пространства знаний некоторой предметной области, то рассматриваемые в этом разделе позволяют вести описание экземпляров классов ***InformationObject и Concept*** на третьем уровне в соответствии с введенной моделью на втором уровне.

Каждому классу атрибутов ***ResourceAttribute*** и ***ThesaurusAttribute*** соответствует класс ***ObjectAttributeValue*** и ***ConceptAttributeValue*** в котором хранится значение этого атрибута. Это издержки адаптивной модели, в которой для поддержки гибкого описания структуры ресурсов с определением типов атрибутов надо вводить отдельные классы для хранения значений этих атрибутов. В дальнейшем это позволяет организовать атрибутный настраиваемый поиск информационных объектов и упрощает анализ этих значений в системе.

1.1.1. ***ObjectAttributeValue*** – класс экземпляры которого хранят значения конкретного информационного объекта

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithoutURI)

Название свойства	**Тип значений**
Объект (lb:object)	(lb:InformationObject)[1..1]
Значение (lb:value)	Строка (xsd:string) [1..1]
Атрибут (lb:resourceAttribute)	(lb:ResourceAttribute) [1..1]

1.1.2. ***ConceptAttributeValue***– класс экземпляры которого хранят дополнительные значения конкретного концепта тезауруса

Суперклассы: Уникально-идентифицируемый объект (lb:ClassWithoutURI)

Название свойства	**Тип значений**
Концепт (lb:concept)	(lb:Concept) [1..1]
Значение (lb:value)	Строка (xsd:string) [1..1]
Атрибут (lb:thesaurusAttribute)	(lb:ThesaurusAttribute) [1..1]

Для описания экземпляров информационных ресурсов, а именно экземпляров класса ***InformationObject*** используется следующий набор классов ответственных за хранение и предоставление непосредственно данных, являющихся подклассами класса ***ObjectAttributeValue***:

- ***ObjectAttributeFileValue*** – класс значений атрибутов, соответствующих некоторому информационному объекту, ответственных за хранение данных в виде файла

- ***ObjectAttributeHrefValue*** – класс значений атрибутов, соответствующих некоторому информационному объекту, ответственных за хранение данных в виде ссылки
- ***ObjectAttributeNumberValue*** – класс значений атрибутов, соответствующих некоторому информационному объекту, ответственных за хранение данных в виде числа
- ***ObjectAttributeDateValue*** – класс значений атрибутов, соответствующих некоторому информационному объекту, ответственных за хранение данных в виде даты
- ***ObjectAttributeObjectValue*** – класс значений атрибутов, соответствующих некоторому информационному объекту, ответственных за хранение данных в виде связи с некоторым информационным объектом входящим в контент библиотеки определенного типа
- ***ObjectAttributeStringValue*** – класс значений атрибутов, соответствующих некоторому информационному объекту, ответственных за хранение данных в виде строки
- ***ObjectAttributeTaxonomyValue*** – класс значений атрибутов, соответствующих некоторому информационному объекту, ответственных за хранение данных в виде связи с некоторым таксоном определенного словаря или классификатора
- ***ObjectAttributeTextValue*** – класс значений атрибутов, соответствующих некоторому информационному объекту, ответственных за хранение данных в виде текста
- ***ObjectAttributeXMLValue*** – класс значений атрибутов, соответствующих некоторому информационному объекту, ответственных за хранение данных в виде данных в xml синтаксисе

Эти классы являются подклассами суперкласса ***ObjectAttributeValue***, который связан с классом ***ResourceAttribute*** отношением по принципу *часть – целое* посредством свойства ***attribute***. Для этого свойства определены подсвойства:

- ***attributeXML*** связывающие классы ***ObjectAttributeXMLValue*** ***ResourceAttributeXML***
- ***attributeText*** связывающие классы ***ObjectAttributeTextValue*** ***ResourceAttributeText***
- ***attributeTaxonomy*** связывающие классы ***ObjectAttributeTaxonomyValue*** ***ResourceAttributeTaxonomy***
- ***attributeString*** связывающие классы ***ObjectAttributeStringValue*** ***ResourceAttributeString***
- ***attributeObject*** связывающие классы ***ObjectAttributeObjectValue*** ***ResourceObjectAttribute***
- ***attributeDate*** связывающие классы ***ObjectAttributeDateValue*** ***ResourceDateAttribute***
- ***attributeNumber*** связывающие классы ***ObjectAttributeNumberValue*** ***ResourceNumberAttribute***
- ***attributeHref*** связывающие классы ***ObjectAttributeHrefValue*** ***ResourceHrefAttribute***
- ***attributeFile*** связывающие классы ***ObjectAttributeFileValue*** ***ResourceFileAttribute***

Для класса ***InformationObject*** определено многозначное свойство ***attributeValues***, которое содержит множество экземпляров класса ***ObjectAttributeValue.*** Схема связей приведена на рисунке 1.

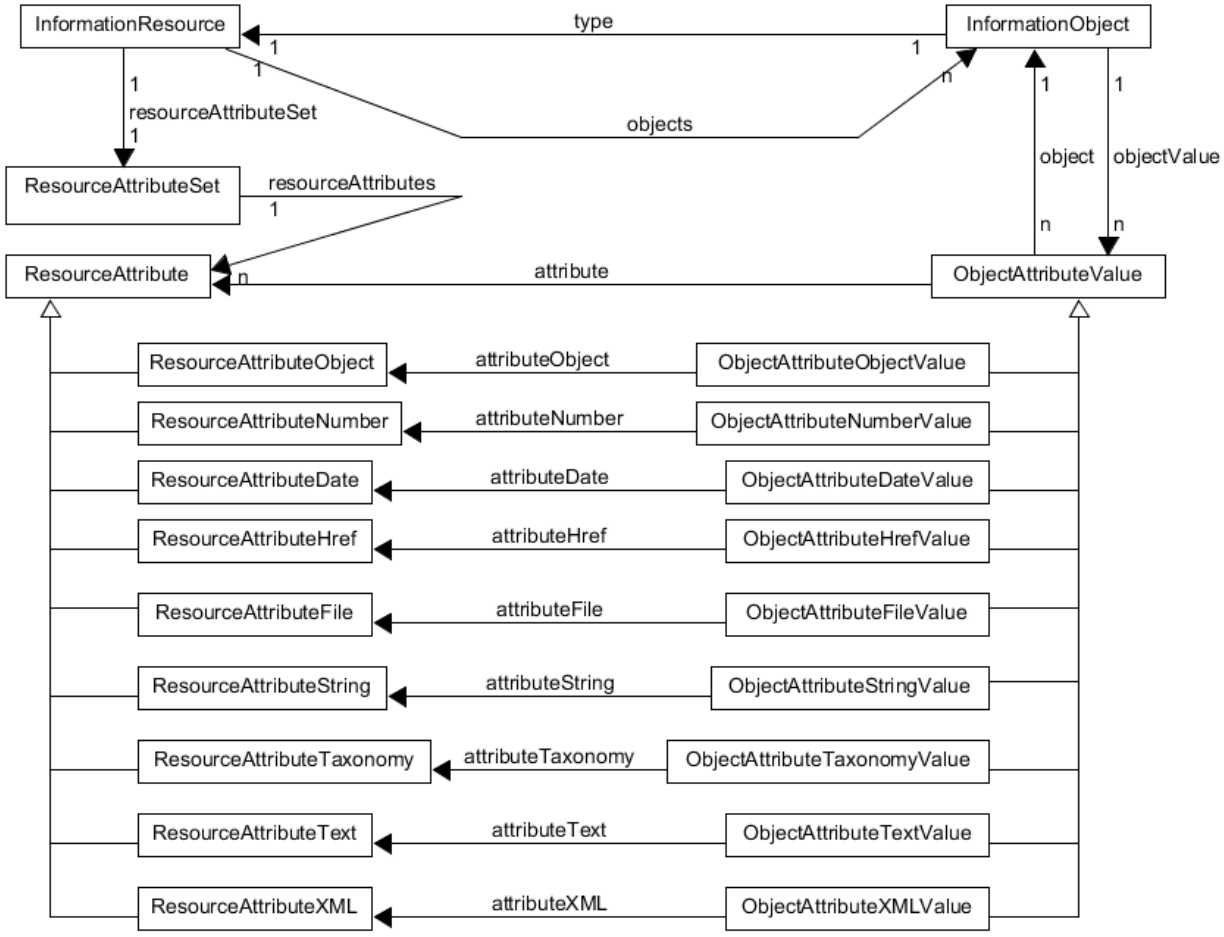

Рисунок 1

Для описания расширенного представления понятий тезауруса, а именно экземпляров класса ***Concept*** используется следующий набор классов ответственных за хранение и предоставление непосредственно данных, являющихся подклассами класса ***ConceptAttributeValue***:

- ***ConceptAttributeFileValue*** – класс значений атрибутов, соответствующих концепту некоторого тезауруса, ответственных за хранение данных в виде файла
- ***ConceptAttributeHrefValue*** – класс значений атрибутов, соответствующих концепту некоторого тезауруса, ответственных за хранение данных в виде ссылки
- ***ConceptAttributeNumberValue*** – класс значений атрибутов, соответствующих концепту некоторого тезауруса, ответственных за хранение данных в виде числа

- ***ConceptAttributeDateValue*** – класс значений атрибутов, соответствующих концепту некоторого тезауруса, ответственных за хранение данных в виде даты
- ***ConceptAttributeObjectValue*** – класс значений атрибутов, соответствующих концепту некоторого тезауруса, ответственных за хранение данных в виде связи с некоторым информационным объектом входящим в контент библиотеки определенного типа
- ***ConceptAttributeStringValue*** – класс значений атрибутов, соответствующих концепту некоторого тезауруса, ответственных за хранение данных в виде строки
- ***ConceptAttributeTaxonomyValue*** – класс значений атрибутов, соответствующих концепту некоторого тезауруса, ответственных за хранение данных в виде связи с некоторым таксоном определенного словаря или классификатора
- ***ConceptAttributeTextValue*** – класс значений атрибутов, соответствующих концепту некоторого тезауруса, ответственных за хранение данных в виде текста
- ***ConceptAttributeXMLValue*** – класс значений атрибутов, соответствующих концепту некоторого тезауруса, ответственных за хранение данных в виде данных в xml синтаксисе

Эти классы являются подклассами суперкласса ***ConceptAttributeValue***, который связан с классом ***ThesaurusAttribute*** отношением по принципу *часть – целое* посредством свойства ***attribute***. Для этого свойства определены подсвойства:

- ***attributeXML*** связывающие классы ***ConceptAttributeXMLValue*** ***ThesaurusAttributeXML***
- ***attributeText*** связывающие классы ***ConceptAttributeTextValue*** ***ThesaurusAttributeText***
- ***attributeTaxonomy*** связывающие классы ***ConceptAttributeTaxonomyValue*** ***ThesaurusAttributeTaxonomy***

- ***attributeString*** связывающие классы ***ConceptAttributeStringValue ThesaurusAttributeString***
- ***attributeObject*** связывающие классы ***ConceptAttributeObjectValue ThesaurusObjectAttribute***
- ***attributeDate*** связывающие классы ***ConceptAttributeDateValue ThesaurusDateAttribute***
- ***attributeNumber*** связывающие классы ***ConceptAttributeNumberValue ThesaurusNumberAttribute***
- ***attributeHref*** связывающие классы ***ConceptAttributeHrefValue ThesaurusHrefAttribute***
- ***attributeFile*** связывающие классы ***ConceptAttributeFileValue ThesaurusFileAttribute***

Для класса ***Concept*** определено многозначное свойство ***conceptValues***, которое содержит множество экземпляров класса ***ConceptAttributeValue.*** Схема связей приведена на рисунке 1.

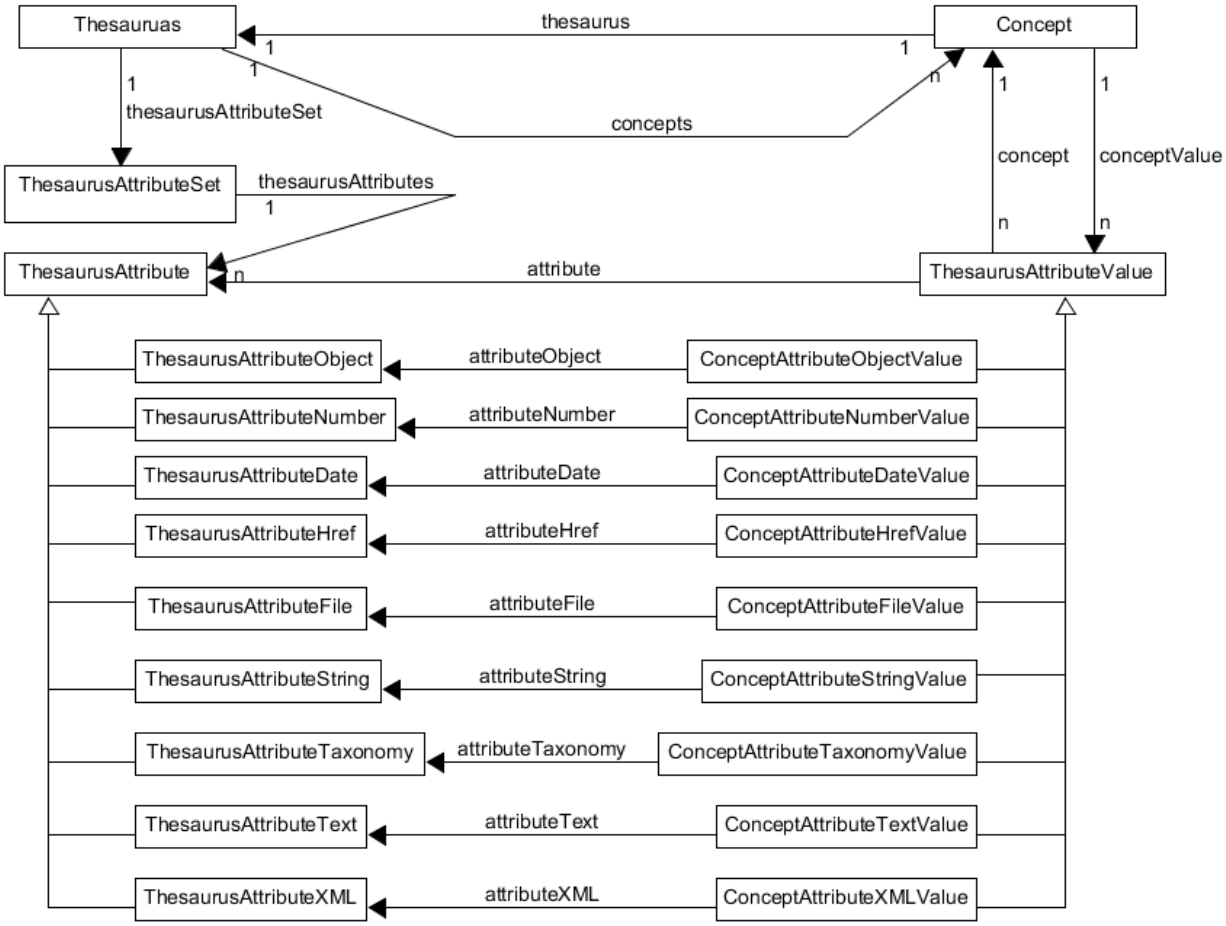

Рисунок 2

4.2. Правила вывода

Математический аппарат, лежащий в основе дескриптивных логик, на которых базируются онтологии, предоставляет средства логического вывода новых фактов на основе имеющихся. Логический вывод позволяет выявлять скрытые знания и находить противоречия в онтологии.

В онтологии обычно правила для неявных связей задаются с помощью правил SWRL [44, 157]. SWRL как расширение OWL помогает описать *абстрактный* механизм оперирования объектами предметной области и ее закономерности. SWRL дает возможность выводить новые факты из существующих утверждений, что повышает эффективность описания предметной области. На базе введенных понятий при генерации онтологии предметной области правила для конкретной предметной области генерируются в синтаксисе SWRL.

Применение правил имеет ряд преимуществ. Рассмотренные выше классы онтологии и отношения между ними представляют собой факты или знания о предметной области. Занесение всех знаний или фактов 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*.

Для обеспечения безопасности правил следует использовать правила безопасности, которые используют только те переменные, которые могут быть привязаны к конкретному классу онтологии.

4.3. Выводы

В этой главе были рассмотрены основные понятия онтологии семантической библиотеки и научного пространства знаний и связи между ними. Был предложен набор основных понятий для построения описания произвольной предметной области.

.

5. Архитектура семантической библиотеки

Рассмотрим 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 использование. При этом каждая подсистема может состоять из нескольких уровней абстракции, и каждый уровень может иметь свою архитектуру.

На основе предложенного метода построения семантической библиотеки для некоторой области научного пространства разработана информационная система LibMeta, в рамках которой задается описание предметной области с терминологической поддержкой и с возможностью интеграции данных из разных источников данных *почти* удовлетворяющим требованиям, предъявляемым к источникам данных в LOD. Слово *почти*, в данном контексте означает, что возможно, одно требование, касающееся связанности данных с другими источниками может не выполняться, но с помощью разработанной системы, появляется возможность достаточно просто выполнить его. Для этого от пользователя – эксперта в предметной области не требуется специальных технических знаний об используемом для этого стеке технологий LOD [126].

Рассмотрим набор подсистем разработанной информационной системы и их взаимодействие. Каждая из этих подсистем отвечает за определенную функциональность и использует определенное подмножество понятий из информационной модели.

5.1. Основная функциональность LibMeta

Основная функциональность LibMeta:

- создание/просмотр/редактирование информационных ресурсов и их структуры;
- создание/просмотр/редактирование информационных объектов и их структуры;
- подключение источников данных;
- загрузка данных из подключенных источников данных, в дальнейшем становящихся частью контента библиотеки;
- создание/просмотр/редактирование структуры тезауруса поддерживаемой предметной области;
- создание/просмотр/редактирование понятий тезауруса
- пакетная загрузка данных составляющих контент библиотеки;
- атрибутный/семантический/полнотекстовый поиск и навигация по доступным информационным объектам системы;
- атрибутный/семантический/полнотекстовый поиск по источникам данных;
- создание/просмотр/редактирование коллекций информационных объектов;
- формирование онтологии предметной области по описанию структуры информационных ресурсов и тезауруса;
- предоставление данных составляющих контент системы в машиночитаемом формате;
- выделение связей между информационными объектами и понятиями тезауруса;
- поддержка семантических меток или *фолксономии* [78, 79, 80] для описания тематической направленности информационных объектов;
- создание/просмотр/редактирование области интересов пользователя;
- создание рекомендательной системы:
    a. на основе описания интересов пользователя;
    b. на основе рассматриваемого тезауруса предметной области;
- поддержка микротезаурусов пользователей на основе тезауруса предметной области.

Функциональность LibMeta, доступная для всех публичных пользователей:

- просмотр информационных ресурсов и их структуры;
- просмотр информационных объектов и их структуры;
- атрибутный/семантический/полнотекстовый поиск и навигация по доступным ресурсам системы;
- атрибутный и семантический поиск по источникам данных;
- просмотр общедоступных коллекций информационных объектов.

С точки зрения авторизированного пользователя, семантическая библиотека обеспечивает ему дополнительно следующую функциональность:

- определение своего микротезауруса как расширение некоторого узла определенного в системе основного терминологического тезауруса. Также обеспечивается поддержка создания так называемых *аннотационных онтологий* или *онтологий пользователей* (фолксономии), которые представляют собой коллективный словарь пользователей, составленный в результате процесса проставления семантических меток ими для ресурсов;
- определение собственных коллекций информационных объектов;
- организация совместных тематических коллекций для групп пользователей;
- атрибутный и семантический поиск по источникам данных с возможностью сохранения результатов поиска;
- пользователь в роли администратора системы имеет доступ ко всей вышеопределенной функциональности и может воспользоваться дополнительной, доступной только ему функциональностью:
    a. может по запросу пользователей расширять описания типов ресурсов или создавать новые;
    b. может по запросу пользователей включать их объекты ресурсов в общедоступный список объектов;
    c. для групп пользователей делать доступными возможности редактирования определенных типов ресурсов или таксономий;
    d. редактировать группы и роли пользователей и набор доступных им операций;

e. осуществлять редактирование и настройку основного терминологического тезауруса и его связей;
f. добавлять источники данных.

5.2. Подсистема описания контента информационной системы

5.2.1.Основные понятия

За универсальность определения контента системы отвечает набор понятий, составляющих информационную модель контента библиотеки Libmeta: *информационный ресурс* и *информационный объект*, которые описывают экземпляры ресурсов. *Информационный ресурс,* является основной единицей описания контента библиотеки, а *информационный объект* представляет экземпляры информационных ресурсов. Каждый из них имеет собственный уникальный идентификатор, в соответствии с требованиями *LOD*. Фактически семантическое значение *информационного ресурса* является эквивалентным понятию *класса онтологии* с некоторыми ограничениями в его описании. Структура описания информационных объектов определяется понятиями *атрибут* и *набор атрибутов*, которые определяются при описании соответствующего ресурса. Атрибут является элементом описания свойства ресурса, а набор атрибутов определяется как коллекция атрибутов разных видов. Типы атрибутов следующие: *атрибут, файловый, объектный, числовой, текстовый, строковый*. При подключении подсистемы управления таксономиями появляется новый вид атрибута - *таксономический*, который будет описан ниже в соответствующем разделе. Помимо определения круга значений атрибута, важной характеристикой являются его тип и определение количества его значений. Для описания конкретного информационного ресурса используется понятие *значение атрибута*, которое тесно связано с понятием *атрибут* и является фактически контейнером для хранения конкретных значений *информационного объекта* определенного типа.

Эти понятия обеспечивают структурированное описание контента и обеспечивают поддержку его адаптируемости. Такой подход также обеспечивает

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

В общем случае, конкретная реализация модели контента библиотеки может быть основана на некоторой импортируемой онтологии, классы которой превращаются в ресурсы, свойства описываются в терминах атрибутов Libmeta, наборы атрибутов определяют фактически домены свойств онтологий. При построении модели ресурсов библиотеки на основе этой онтологии сохраняются все URI свойств, отношений и классов выбранной онтологии. При необходимости при импортировании выбранной онтологии в систему можно изменить набор понятий, расширив или наоборот сократив его средствами системы.

Конечно, такой способ отображения онтологии на понятия системы LibMeta не сохраняет весь возможный перечень ограничений, накладываемых на свойства и классы онтологии изначально, но структурная ее часть сохраняется, что является достаточным для решения задач, определенных в рамках системы.

На рисунке 3 приведены основные понятия, используемые для конструирования описания предметной области в рамках этой подсистемы.

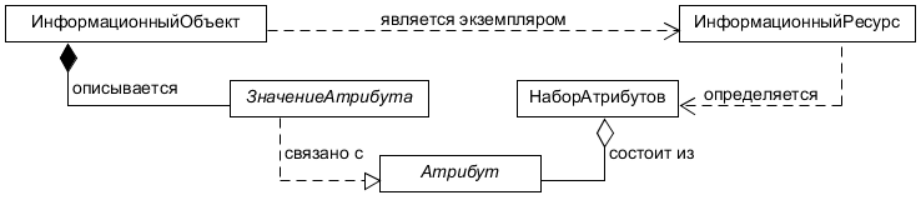

Рисунок 3

При описании *информационных ресурсов* и определении *набора* их *атрибутов* важную роль играют *виды атрибутов*, которые формируют структурное описание *ресурса*. Атрибуты делятся на несколько пересекающихся видов: *поисковые*, *описательные*, *административные*, *идентифицирующие*. При формировании интерфейсов поиска важную роль играют именно поисковые атрибуты, которые используются при выполнении атрибутного поиска по типам ресурсов. Результатом такого поиска являются объекты, краткое описание которых представлено пользователю посредством описательных атрибутов.

Фактически в рамках этой подсистемы выполняется первичная настройка конфигурации контента библиотеки и ее интерфейсов под конкретную предметную область. На рисунке 4 изображена последовательность действий пользователя по настройке системы. Строго говоря, три блока сверху можно

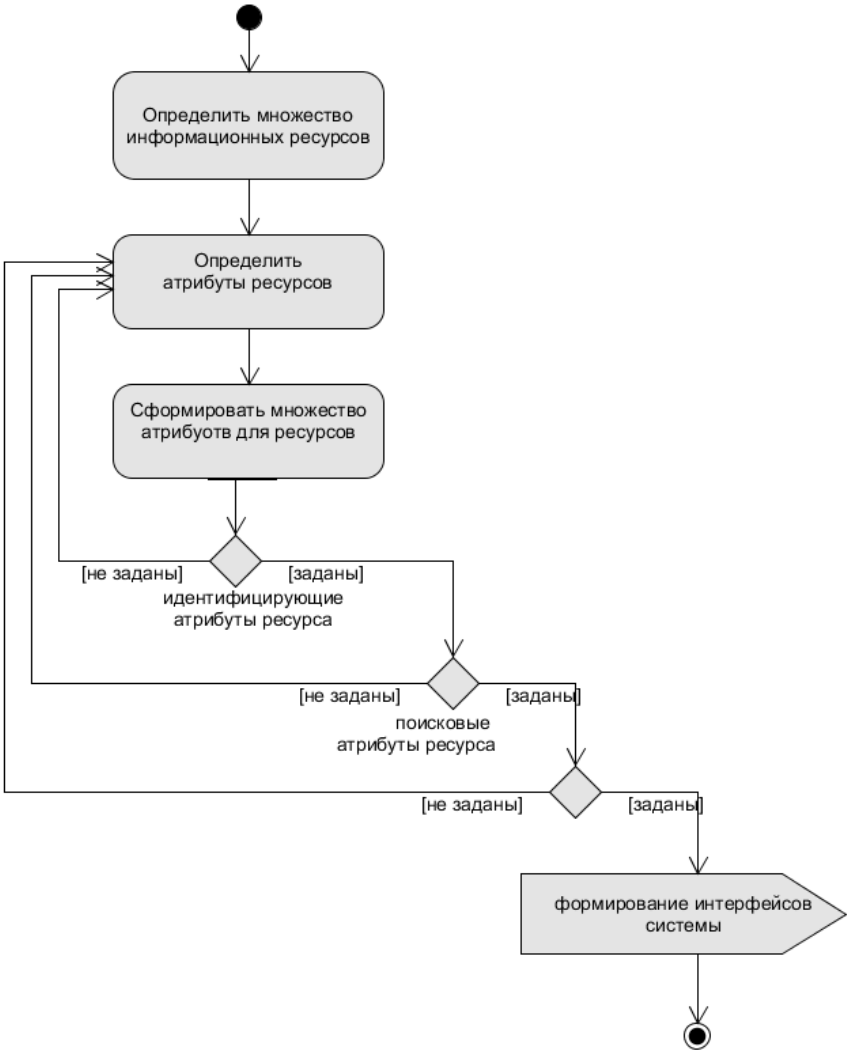

Рисунок 4

назвать отдельными поддеятельностями, поскольку определение информационных ресурсов, их атрибутов и множества атрибутов требует нескольких действий, но для упрощения картины эти действия были объединены каждый в один блок.

5.2.2.Поддерживаемая функциональность.

- Создание информационных ресурсов;
- Создание атрибутов информационных ресурсов;
- Создание набора атрибутов для определения структуры некоторого информационного ресурса;
- Формирование интерфейсов просмотра/создания/редактирования для информационных объектов на основе описания структуры информационных ресурсов;
- Формирование интерфейсов поиска информационных объектов на основе описания структуры информационных ресурсов;

5.3. Подсистема управления тезаурусом

5.3.1.Основные понятия

Рассмотрим базовые понятия, используемые в подсистеме управления тезаурусом предметной области. Эти понятия позволяют определить также совокупность словарей и классификаторов, являющихся частью тезауруса, а также разнообразные связи между их понятиями.

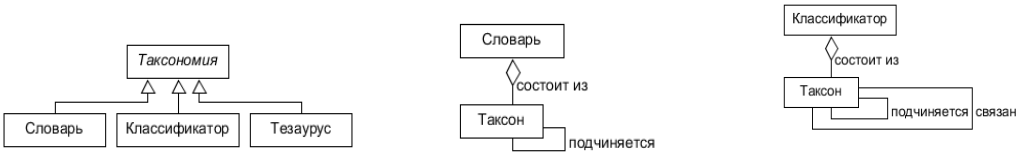

Рисунок 5

Для этого используются понятия: *таксон* и *таксономия*. Таксон представляет собой элемент таксономии с определенным набором свойств, необходимым для его базового представления, а таксономия определяет набор доступных связей между составляющими таксономию таксонами и ресурсами системы. Для описания дополнительных связей между таксонами вводятся отношения между таксонами, которые позволяют определять и описывать новые

связи в рамках информационной системы. На рисунке 5 представлены поддерживаемые подпонятия таксономии согласно определению семантической библиотеки: *словарь*, *классификатор* и *тезаурус* и отображаются используемые по умолчанию связи в таксономиях между определяющими их таксонами. По умолчанию в системе доступно только два типа связей между таксонами: иерархическая и нетипизированная горизонтальная.

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 6.

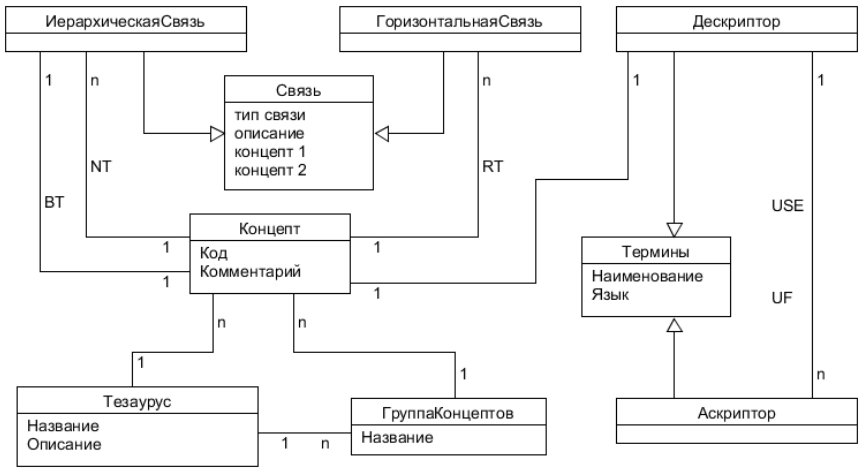

Рисунок 6

Для тезауруса возможно доопределить атрибуты в соответствии с моделью данных. Например, для определения связи концептов с информационными объектами используется понятие *объектного атрибута*. Это отношение обеспечивает возможность подключения к любому типу ресурсов в процессе жизнедеятельности системы. Такой подход с одной стороны позволяет избежать избыточности на начальном этапе проектирования системы, с другой стороны позволяет обеспечить представление практически любых связей.

В рамках этой подсистемы выполняется настройка конфигурации тезауруса библиотеки и ее интерфейсов под конкретную предметную область, путем ее терминологического ограничения и определения явных связей с контентом библиотеки. На рисунке 7 изображена последовательность действий пользователя по настройке тезауруса. Строго говоря, блоки слева можно назвать отдельными поддеятельностями, поскольку задание дополнительных атрибутов тезауруса требует нескольких действий, но для упрощения картины эти действия были объединены в один блок.

Рисунок 7

5.3.2.Основная функциональность.

- Создание/просмотр/редактирование тезауруса;
- Создание/просмотр/редактирование атрибутов расширяющих структуру понятия тезауруса;
- Создание/просмотр/редактирование связей между понятиями тезауруса;
- Создание/просмотр/редактирование связей между понятиями тезауруса и информационными объектами;
- Формирование интерфейсов просмотра/создания/редактирования понятий тезауруса с учетов атрибутов, расширяющих структуру понятия тезауруса;

- Формирование интерфейсов поиска понятий тезауруса, с учетом расширяющих описание атрибутов;
- Разметка текстовых данных информационных объектов терминами тезауруса.

5.4. Подсистема поддержки коллекций

5.4.1.Основные понятия

Для возможности ведения разнообразных коллекций объектов используется понятие *коллекция информационных объектов*, 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.

Коллекции делятся на *типизированные* и *не типизированные*. К первым относятся коллекции, которые определяются на основе таксономии, фактически разбивая объекты коллекции по направлениям, согласно составу таксономии. Ко вторым будем относить произвольные наборы информационных объектов, составленных пользователями, в рамках своего информационного пространства в системе. Ко второму типу также относятся наборы объектов объединенных в результате развития *фолксономии* (наборы семантических меток) в системе.

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 системе поддерживаются все типы коллекций описанных в онтологии:

- Типизированные коллекции

a. Коллекции на основе тезауруса,
b. Коллекции на основе словаря
c. Коллекции на основе классификатора

- Не типизированные коллекции
  a. Произвольная коллекция информационных объектов любых типов
  b. Коллекции на основе тегов

Типизированные коллекции отличаются от не типизированных тем что при их определении явно указывается тип ресурсов для объектов, которые будут включаться в коллекцию.

Для включения информационного объекта в ту или иную коллекцию, в интерфейсы редактирования автоматически добавляются соответствующие элементы для включения/удаления/просмотра.

5.4.2.Основная функциональность

- Создание/редактирование/просмотр метаинформации о коллекции
- Формирование интерфейсов просмотра/создания/редактирования содержимого коллекций;
- Формирование интерфейсов поиска по коллекции;

5.5. Подсистема автоматизированной обработки и представления данных

5.5.1.Основные понятия

Одной из самых распространенных операций является импорт или загрузка данных в библиотеку. Подсистема импорта данных 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. На первом уровне элементы представляют собой сериализованные RDF-ресурсы, и названием элемента является имя RDFS-класса информационного объекта (*lbm:InformationObject*). Описания RDF ресурсов, как первого уровня, так и вложенные, идентифицируется URI с помощью атрибута *rdf:about*. Для каждого объекта первый дочерний элемент указывает тип ресурса набор атрибутов которого определяет описание информационного объекта. Остальные дочерние элементы соответствуют атрибутам этого ресурса. Если свойство имеет несколько значений, для каждого значения указывается свой элемент. Если некоторое значение свойства - примитивного типа данных (строка, число, дата..), то оно указывается в виде текстового содержимого элемента.

На рисунке 8 приведен пример описания информационного объекта подготовленного для загрузки в синтаксисе RDF/XML

Рисунок 8

Данная подсистема также отвечает за предоставление данных системы в машиночитаемом формате по запросу.

5.5.2.Основная функциональность

- Пакетная загрузка данных в машиночитаемом формате (RDF/XML)
- Предоставление информации об информационном ресурсе в машиночитаемом формате (RDF/XML)
- Предоставление информации об информационном объекте в машиночитаемом формате (RDF/XML)
- Предоставление информации об информационных объектах определенного ресурса в машиночитаемом формате (RDF/XML)
- Предоставление онтологии предметной области в машиночитаемом формате (OWL в синтаксисе RDF/XML)

5.6. Подсистема реализации задач интеграции данных из источников LOD

5.6.1.Основные понятия

Для решения задач интеграции данных из источников LOD вводится понятие *источник данных*, которому 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.

Рисунок 9

Если конкретная реализация модели контента библиотеки основана на некоторой импортируемой онтологии и онтология используется в источнике данных, то предусмотрен механизм взаимно однозначного отображения свойств и классов онтологий подключаемого набора данных в термины LibMeta полуавтоматическим способом. Таким образом, мы формируем интеграционный узел, который позволяет устанавливать взаимосвязи с источниками данных, расположенными в LOD. На рисунке 9 приведена схема связей понятия *источника данных* с основными понятиями, определяющими контент библиотеки. Рисунок 10 иллюстрирует взаимодействие пользователя с подсистемой для получения результатов своего запроса.

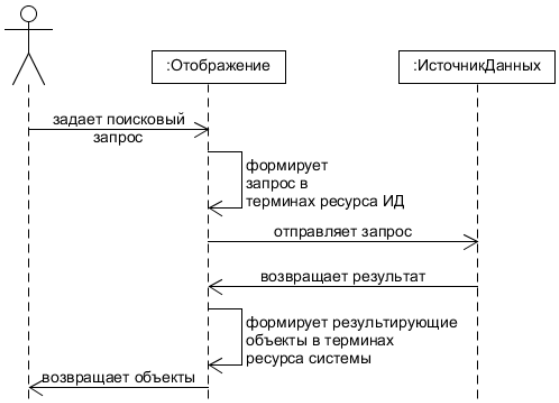

Рисунок 10

В отличие от других систем интеграции данных из LOD, в данном случае не представляет интереса построение иерархий классов при трансляции ресурсов на источник данных, поэтому для проставления связей используется связь, которая указывает, что два разных класса могут иметь одинаковых представителей. Эта связь 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 используется предположение, что онтология источника данных частично совместима со структурой ресурсов семантической библиотеки. Это означает, что хотя бы один ресурс предметной онтологии может быть транслирован в некоторый класс в онтологии источника данных и, таким образом, требуется лишь минимальное частичное соответствие ресурсу LibMeta.

5.6.2.Основная функциональность

- Создание/редактирование/просмотр информации об источнике данных
- Подключение интегрируемых типов ресурсов
- Отображение свойств информационных ресурсов библиотеки на модель данных источника
- Поиск объектов идентичных информационным объектам из библиотеки
- Сохранение результатов

5.7. Подсистема поддержки пользователей LibMeta.

5.7.1.Основные понятия

Важной составляющей любой информационной системы являются ее пользователи. Рассмотрим основные понятия подсистемы поддержки пользователей. Основными понятиями этой подсистемы являются *пользователь, роль, разрешение, информационный ресурс, информационный объект, область интересов*. 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. Структура *области интересов* пользователя формируется из трех типов понятий:

- концепты тезауруса;
- семантические метки;
- таксоны, соответствующие классификаторам/словарям использующимся в системе.

При этом каждый пользователь волен создавать свои коллекции ресурсов в рамках своих интересов, пользуясь доступными средствами соответствующих подсистем, что и иллюстрируется на рисунке 11.

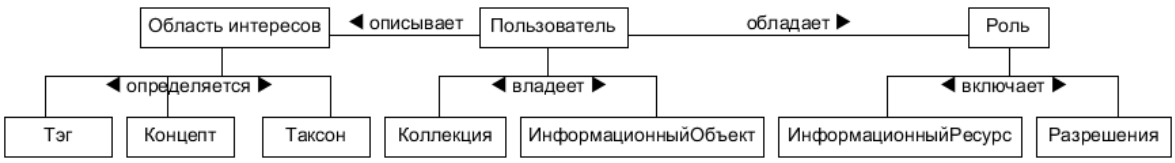

Рисунок 11

5.7.2.Основная функциональность

- Регистрация/авторизация пользователя в системе;

- Поддержка иерархии пользователей системы по наличию прав доступа к работе с информационными ресурсами;
- Поддержка работы пользователя с собственными коллекциями;
- Поддержка сохранения результатов поиска;
- Поддержка работы с собственными информационными ресурсами и объектами;
- Определение области интересов и ее детализация. Поддержка создания своего микротезауруса области интересов.

5.8. Подсистема поддержки микротезауруса пользователя

5.8.1.Основные понятия

Определение структуры микротезауруса 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. Эти ветки описываемые пользователем в рамках своей области интересов не отображаются в основной тезаурус, но могут содержать весь набор связей между концептами определяемых тезаурусом. Концепты микротезауруса пользователь может использовать также для классификации информационных объектов вручную или автоматически, воспользовавшись средствами предоставляемыми подсистемой. Таким образом, следует, что для каждого *микротезауруса* определены его *владелец*, выделены *связи* с информационными объектами и становится возможным определение семантических тегов на основе связи меток и концептов микротезауруса.

Микротезаурусы пользователей – экспертов могут помочь редакторам основного тезауруса принять решение о расширении основного тезауруса на их основе.

5.8.2.Основная функциональность

- Создание/редактирование/просмотр микротезауруса;
- Выделение связей концептов микротезауруса с информационными объектами системы;
- Формирование семантических тегов на основе связанных с концептами микротезауруса информационных объектов и их семантических меток;
- Формирование интерфейсов навигации по связанным семантическим тегам;
- Расширение состава концептов основного тезауруса с учетом микротезаурусов.

5.9. Рекомендательная подсистема

Рассматриваемая подсистема рекомендаций будет основываться как на анализе текстовой информации из метаданных информационных объектов, так и, при наличии, в текстах являющихся содержимым этих объектов. Так же важную роль играют связи, которые связывают концепты тезауруса и информационные объекты.

5.9.1.Поддержка семантических меток. Основные понятия

Набор основных понятий, которые используются в этой подсистеме *объект, семантический тег, понятие тезауруса* невелик, но возможности семантической навигации, предоставляемые этой подсистемой трудно недооценить. *Семантический тег* – это семантическая метка объекта, которая представляет собой ключевое слово соответствующее информационному объекту библиотеки, характеризующее его тематическую направленность. Особенностью семантического тега является наличие связи с концептами тезауруса. Такие теги являются чрезвычайно информативны и полезны для информационных объектов тех ресурсов библиотеки, которые могут сочетать как структурированные, так и неструктурированные данные.

Поддерживается несколько способов сопоставления семантических меток информационным объектам:

- проставление меток пользователями;
- извлечение семантических меток явно из метаданных составляющих описание информационных объектов;
- применение алгоритмов извлечения ключевых слов из текстов и их использование в качестве меток;

Формирование семантического тега осуществляется автоматически при связи меток с концептами тезауруса. Использование связки меток и концептов с учетом связанных информационных объектов позволяет формировать:

- наборы ключевых слов для концептов тезауруса
- образовывать категории или фасеты со счетчиками для результатов поиска, с разной степенью гранулярности;

Помимо того что, использование семантических тегов позволяет улучшить контроль над данными предоставленными в системе, их использование так же позволяет улучшить навигацию и отобразить разные грани семантического представления информации пользователям.

5.9.2.Рекомендации по области интересов. Основные понятия

Дополнительная функциональность рекомендательной подсистемы будет доступна пользователям при описании своей области интересов. Как было сказано выше, структура *области интересов* пользователя формируется из трех типов понятий:

- концепты тезауруса;
- семантические метки;
- таксоны, соответствующие классификаторам/словарям, использующимся в системе.

После 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. На следующем этапе оцениваются связанные концепты тезауруса (по иерархическим и горизонтальным связям) и

формируются категории объектов связанных с этими концептами на основе уточняющих или более общих. Такой же анализ проводится при необходимости для таксонов словарей/классификаторов. Также проводится анализ набора меток соответствующей области интересов и наиболее часто встречающихся с ними попарно. На основе этого также формируются категории объектов для рекомендаций.

5.9.3.Основная функциональность

- образовывать категории или фасеты со счетчиками для составления рекомендаций с учетом концептов из области интересов пользователя.
- подбирать рекомендации меток для информационных объектов по связанным концептам тезауруса;
- подбирать рекомендации подходящих для информационных объектов концептов тезауруса по семантическим меткам;
- подбор рекомендации подходящих информационных объектов по концептам тезауруса из области интересов
- подбор рекомендации подходящих информационных объектов по семантическим меткам из области интересов
- подбор рекомендации подходящих информационных объектов по элементам классификаторов/словарей из области интересов

5.10. Выводы

В этой главе представлен основной набор подсистем необходимых для реализации функциональности семантической библиотеки для некоторой предметной области. Таким образом, получаем возможность реализации главной задачи библиотеки – *семантического/интеллектуального* конструирования научного пространства знаний для некоторой предметной области. То есть наделение его семантикой за счет выделения явно интеллектуально значимых связей, поддержкой семантической разметки. Основным инструментов

конструирования является, конечно, онтология предметной области, которая позволяет осмысленно структурировать и обеспечить связность между ресурсами которые включены в научное пространство знаний предметной области и использование унифицированной терминологической поддержки в виде тезауруса этой предметной области. Для реализации функций открытости научного пространства знаний были реализованы возможности интеграции других источников данных и возможности связывания с их данными. Предоставление функциональности для совместной работы над развитием пространства научного знания, повышает эффективность проводимых в нем исследований и расширяет возможности по его поддержке в актуальном состоянии, несмотря на лавинообразный рост информации последние десятилетия.

6. Программная реализация семантической библиотеки LibMeta

На основе модели понятий, описанной в работе, а также идей Semantic Web и Linked Open Data, была разработана *персональная открытая семантическая цифровая* библиотека LibMeta с системой поддержки работы пользователей с цифровыми ресурсами библиотек и их коллекциями для некоторой научной предметной области, ограниченной терминологически с помощью тезауруса .При реализации LibMeta авторы руководствовались набором основных задач, которые должна решать разрабатываемая система:

1) библиотека должна поддерживать возможность использования медийных объектов или ссылки на них при описании своих объектов, включая текст, аудио-, видеофайлы или любую их комбинацию. Это требование отражается в названии словом *«цифровая»*;

2) типы используемых ресурсов и связи между ними должны быть описаны средствами системы в рамках понятий, составляющих семантическое описание ресурсов контента библиотеки. При этом согласно принципам LOD при описании ресурсов поддерживается использование классов и свойств, ранее используемых онтологий в сообществе, поддерживающем LOD. Эта поддержка выражается либо в непосредственном использовании готовых онтологий при описании ресурсов и связей между ними, либо возможностью ссылок на их элементы, используя связи на уровне описания ресурсов. Это требование отражается в названии словом *«семантическая»*;

3) библиотека должна служить интеграционным узлом, предоставляя возможность связывания своих данных с данными из разных источников, которые включены в облако LOD. Должна также обеспечиваться возможность извлекать данные этой библиотеки в машиночитаемом формате. Это требование отражается в названии словом *«открытая»*;

4) пользователи библиотеки должны иметь возможность организовывать свои коллекции по интересующему их научному направлению, добавляя новые

термины в предметный тезаурус, уточняя, таким образом, область своих интересов. Пользователи должны также иметь возможность осуществлять поиск не только среди объектов в рамках системы, но и по источникам данных без необходимости использования специализированного языка для поисковых запросов. Это требование отражается в названии словом *«персональная»*.

Основные требования, предъявляемые при этом к контенту системы, – *универсальность, структурированность, адаптируемость*, не противоречат перечисленным свойствам и обеспечивают поддержку настраиваемого хранилища метаданных для объектов и расширяемый набор информационных ресурсов. *Универсальность* обеспечивает описание типов ее ресурсов и объектов независимо от предметной области и области интересов пользователей. *Структурированность* описания обеспечивает поддержку связей между различными типами ресурсов как внутри системы, так и вне ее, исходя из определений LOD. *Адаптируемость* описания ресурсов обеспечивает возможность добавления новых свойств и связей в процессе развития системы и обеспечивает настройку пользовательских интерфейсов под эти изменения [143].

Фактически LibMeta предоставляет функциональность конструирования пространства научного знания предметной области в рамках библиотеки согласно перечисленным требованиям и на начальном этапе при установке системы требуется всего лишь произвести настройку системы под конкретную предметную область, описав ее ресурсы и таксономии, которые будут очерчивать тематически предметную область ее ресурсов и таким образом составлять ее тезаурус.

Настоящая глава посвящена описанию и анализу прототипа системы, созданной на основе предложенному в диссертации подходу. Описываются возможности прототипа и приводятся результаты его исследования на соответствие предъявляемым к системе требованиям, которые сформулированы в разделе 1.6 (постановка задачи).

6.1. Особенности программной реализации

Программный код прототипа системы LibMeta составляет около 20 тысяч строк без учета сторонних модулей. Основная часть кода написана на языках Groovy, Java, HTML, Javascript и CSS. В таблице представлены результаты анализа кода системы, полученные с помощью инструмента CLOC. В таблицу не включены программные коды многочисленных сторонних модулей, которые используются в системе. При подсчете учитывались только файлы, написанные автором диссертации

Language	Files	Blank	Comment	Code
XML	13	33	0	81184
Grails	187	2447	911	9775
Groovy	135	2090	969	7474
Javascript	2	19	4	73
SUM	*337*	*4589*	*1884*	*98506*

Прототип системы LibMeta реализован с применением следующих технологий:

- 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.

  Прототип системы LibMeta состоит из 6 модулей-приложений, которые были созданы автором диссертации, которые включают 39 сторонних

плагинов входящих как в состав Grails, так и созданных сторонними разработчиками. Отметим, что именно в первых 6 модулях – приложениях учитывается специфика системы, остальные плагины являются вспомогательными.

- Для реализации работы с онтологиями в формате OWL мы использовали в фреймворк Apahe Jena, который предоставляет API для чтения/записи данных из/в RDF графы.
- Для отображения информации используется подсистема шаблонов, которая входит в состав Grails. В системе LibMeta применяются шаблоны с использованием технологии GSP, аналогичной технологии JSP, позволяющей разработчикам создавать содержимое, которое имеет как статические, так и динамические компоненты. Страница GSP содержит текст двух типов: статические исходные данные, которые могут быть оформлены в одном из текстовых форматов HTML или XML, и GSP-элементы, которые конструируют динамическое содержимое. Кроме этого могут использоваться библиотеки GSP-тегов, а также EL (Expression Language), для внедрения Groovy-кода в статичное содержимое GSP-страниц.
- Для оформления внешнего вида выходных HTML-документов используется язык CSS.
- В качестве языка сценариев, исполняемых на стороне клиента, в системе используется язык Javascript. В системе применяется широко распространенный Javascript-фреймворк jQuery, набор библиотек для разработки пользовательского интерфейса jQuery UI, а также сторонние модули к jQuery. Фреймворк jQuery распространяется в открытых исходных кодах по лицензиям MIT и GPL.
- В качестве веб-сервера используется Apache Tomcat — контейнер сервлетов с открытым исходным кодом, разрабатываемый Apache Software Foundation. Реализует спецификацию сервлетов, написан на языке Java. Tomcat позволяет запускать веб-приложения, содержит ряд программ для

самоконфигурирования и используется в качестве самостоятельного веб-сервера или в качестве сервера контента в сочетании с веб-сервером Apache. Распространяется в открытых исходных кодах по лицензии Apache License 2.0.

- В качестве системы управления базами данных (СУБД) в комплексе LibMeta применяется продукт PostgreSQL9. Следует подчеркнуть, что объектно-реляционное отображение (ORM), которое входит в состав Grails, позволяет использовать различные системы управления базами данных без изменения Groovy-кода основного программного комплекса. Это обстоятельство существенно облегчает смену используемой СУБД. В число поддерживаемых СУБД входят SQL Server, MySQL, SQLite. Отметим, что MySQL, SQLite и PostgreSQL являются системами с открытым исходным кодом и распространяются по лицензиям, совместимым с GPL.
- Для реализации отображения ORM на СУБД применяется библиотека Hibernate. Распространяется свободно на условиях GNU Lesser General Public License.

6.2. Практическая апробации

6.2.1. Семантическая библиотека «Обыкновенные дифференциальные уравнения»

Рассмотрим в качестве примера реализации семантической библиотеки, на основе изложенной в работе модели, предметную область обыкновенных дифференциальных уравнений (далее ОДУ). На основе предложенной модели было выполнено конструирование библиотеки для этой области. В качестве тезауруса использован тезаурус ОДУ, разработанный коллективом специалистов в этой области.

В данной задаче в качестве внешнего источника данных была использована система MathNet, для выявления дополнительных связей между информационными объектами, а именно персонами и публикациями. В качестве

источника данных было использовано внутреннее RDF хранилище с данными из MathNet.

Структура тезауруса ОДУ

Особенность этого тезауруса заключается в том, что он содержит не только сами понятия и термины, но и ссылки на публикации, в которых вводятся/определяются эти понятия, их математические записи.

Тезаурус «Обыкновенные дифференциальные уравнения» соответствует стандарту ISO, но имеет несколько особенностей, описанных ниже. Понятия в тезаурусе «Обыкновенные дифференциальные уравнения» (в дальнейшем ОДУ) подразделяются на 4 группы:

- ОДУ и системы ОДУ (мнемоническое отображение в тезаурусе: «DE»);
- решение ОДУ и систем ОД (мнемоническое отображение в тезаурусе: «RDE»);
- методы решения ОДУ и систем ОДУ (мнемоническое отображение в тезаурусе: «DM»);
- условия задач и свойства ОДУ и систем ОДУ (мнемоническое отображение в тезаурусе: «DZ»).

Множество терминов вместе с множеством отношений реализовано в виде дескрипторов, недескрипторов, парадигматических отношений.

Не всякий объект предметной области, который должен быть включен в тезаурус, имеет общепринятое терминологическое обозначение. Поэтому понятия тезауруса ОДУ могут быть следующих видов:

- математическая запись;
- математическая запись и название на естественном языке;
- название на естественном языке;

Термины предметной области ОДУ могут содержаться в тезаурусе в виде дескрипторов (идентификатор имеет префикс DE, DM, RDE или DZ), недескрипторов (идентификатор имеет префикс NOD), синонимов (идентификатор имеет префикс SYN).

Понятие описывается группой терминов, среди которых выделяется главный, называемый дескриптором. Остальные термины, описывающие понятие, являются синонимами его дескриптора. Недескрипторы описаны отдельно от понятий, но могут быть с ними связаны ассоциативно.

Между понятиями, обозначаемыми в тезаурусе дескрипторами, определены разные отношения: парадигматические отношения – аналог отношения RT из стандарта ISO, отношения род-вид – BTG и NTG соответственно, часть-целое – NTP и BTP из стандарта ISO, соответственно.

На рисунке 12 представлены понятия тезауруса, связанные иерархически, и для каждого понятия отображаются его горизонтальные связи.

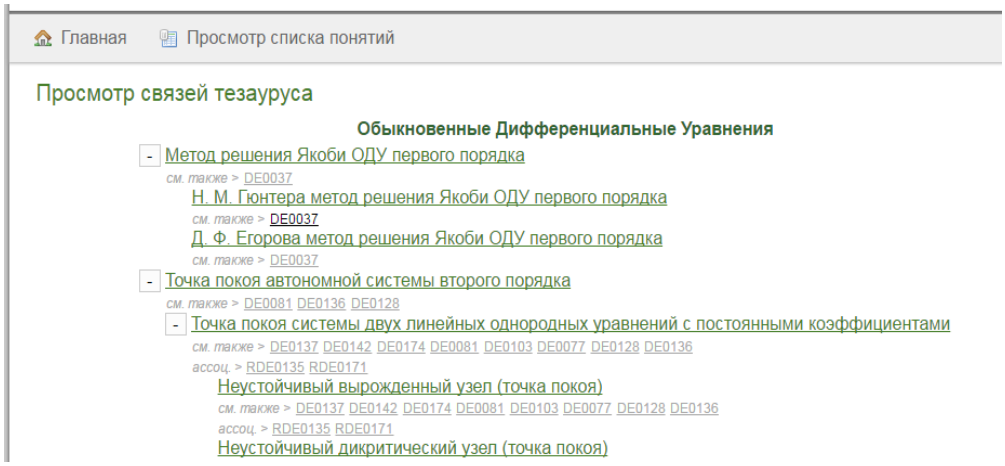

Рисунок 12

Онтология тезауруса ОДУ

Для описания тезауруса ОДУ базовая онтология тезауруса расширяется для того чтобы учесть все особенности модели этого тезауруса. Рассмотрим понятия необходимые для описания на всех уровнях онтологии и связи между ними:

1. На первом уровне используются классы необходимые для описания общей модели, такие как *InformationObject, InformationResource, Thesaurus, Concept, ConceptGroup, Term, Descriptor, Ascriptor, Relation, HierarchicalRelation, FamilyRelation, ThesaurusAttributeSet, ThesaurusAttribute, ThesaurusAttributeValue*.
2. На втором уровне описываются понятия конкретной предметной области как экземпляры, в терминах первого уровня:

a. *Математическая запись* – экземпляр класса *ThesaurusAttribute*, а именно его подкласса *ThesaurusAttributeString*. Используется для хранения строки формулы
b. *Математическое примечание* – экземпляр класса *ThesaurusAttribute*, а именно его подкласса *ThesaurusAttributeText*. Используется для хранения текста с формулами

c. *Литература* – экземпляр класса *InformationResource*, для описания литературы включеной в тезаурус ОДУ
d. *Иерархическая связь Род - Вид* – экземпляр класса *HierarchicalRelation*, для описания литературы включеной в тезаурус ОДУ
e. *Горизонтальная связь Ассоциация* – экземпляр класса *InformationResource*, для описания литературы включеной в тезаурус ОДУ
f. *Горизонтальная связь Синоним* – экземпляр класса *InformationResource*, для описания литературы включеной в тезаурус ОДУ

3. На третьем уровне используем понятия первого и экземпляры второго уровней как определения классов на третьем уровне при заполнении онтологии данными.

Онтология контента ресурсов ОДУ

Объектами библиотеки рассматривались журнальные математические статьи. В качестве примеров типов ресурсов, соответственно, рассматривались *Авторы* и *Публикации*. Был определен набор атрибутов для каждого типа ресурсов в рамках минимального набора свойств на основе Dublin Core для публикаций и FOAF для описания авторов.

Фактически понятия *персоны* и *публикации* представляют собой экземпляры класса *информационный ресурс,* определенного как базовая единица контента семантической библиотеки. Так как каждый ресурс обладает набором атрибутов, для каждого из этих экземпляров задается собственный набор из множества атрибутов, описанных в системе. Множество атрибутов состоит из следующих элементов: *название на языке оригинала, название на русском, фамилия, имя, отчество, электронный адрес, дата рождения, аннотация, идентификатор, автор, деятельность, тип публикации, место рождения, биография, описание, дополнительное заглавие, язык*.

С помощью этого тезауруса был размечен набор публикаций со схожей тематикой. Схожесть тематики публикации тезаурусу ОДУ определялась по ее ключевым словам, соответствующим терминам тезауруса. На рисунке 13 представлен пример связи понятия из ОДУ и найденных публикаций. В качестве связанных объектов могут выступать не только *публикации*, но и, например, *персоны*, в описании деятельности которых могут встречаться соответствующие понятию из ОДУ ключевые слова.

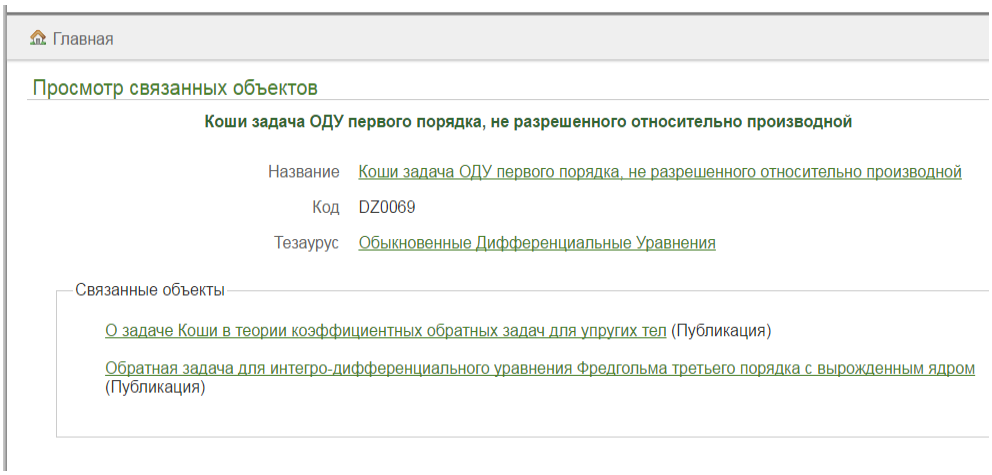

Рисунок 13

Вводится новый тип ресурса *Литература* для тезауруса ОДУ, который фактически представляет собой библиографическое описание источников по которым формировался тезаурус. Этот тип ресурса вводится явно для того чтобы выделить эти публикации в отдельное «неприкасаемое» множество объектов, доступное для редактирования только тем пользователям которые имеют право редактировать термины предметной области определяемыми в тезаурусе.

Экземплярами этого ресурса являются *информационные объекты* системы поэтому в набор понятий контента библиотеки и связей межу понятиями тезауруса и информационными объектами никаких изменений не вносится. Добавляется лишь право редактирования объектов *Литература* только для администраторов предметной области.

Для описания ресурса *литература* используется тот же набор атрибутов, что и для описания ресурса *публикация*. Это позволило отдельно настроить права доступа для всех объектов *литературы*, запретив их модификацию или удаление пользователям, не являющимся редакторами предметной области. В качестве массива данных был предоставлен набор математических статей из «Владикавказского математического журнала» [27] за 19 лет. К настоящему моменту обработаны только библиографические описания статей, а именно *заглавие, аннотация, ключевые слова, авторы, год издания, номер тома, номер выпуска.* Часть публикаций была размечена ключевыми словами, однако термины не разделены между собой и просто перечислялись через запятую в одном поле. Ключевые слова публикации были преобразованы в набор семантических меток соответствующего информационного объекта для каждой извлеченной публикации. В коллекцию публикаций тезауруса ОДУ добавлялись те объекты, в наборе ключевых слов которых находились термины ОДУ.

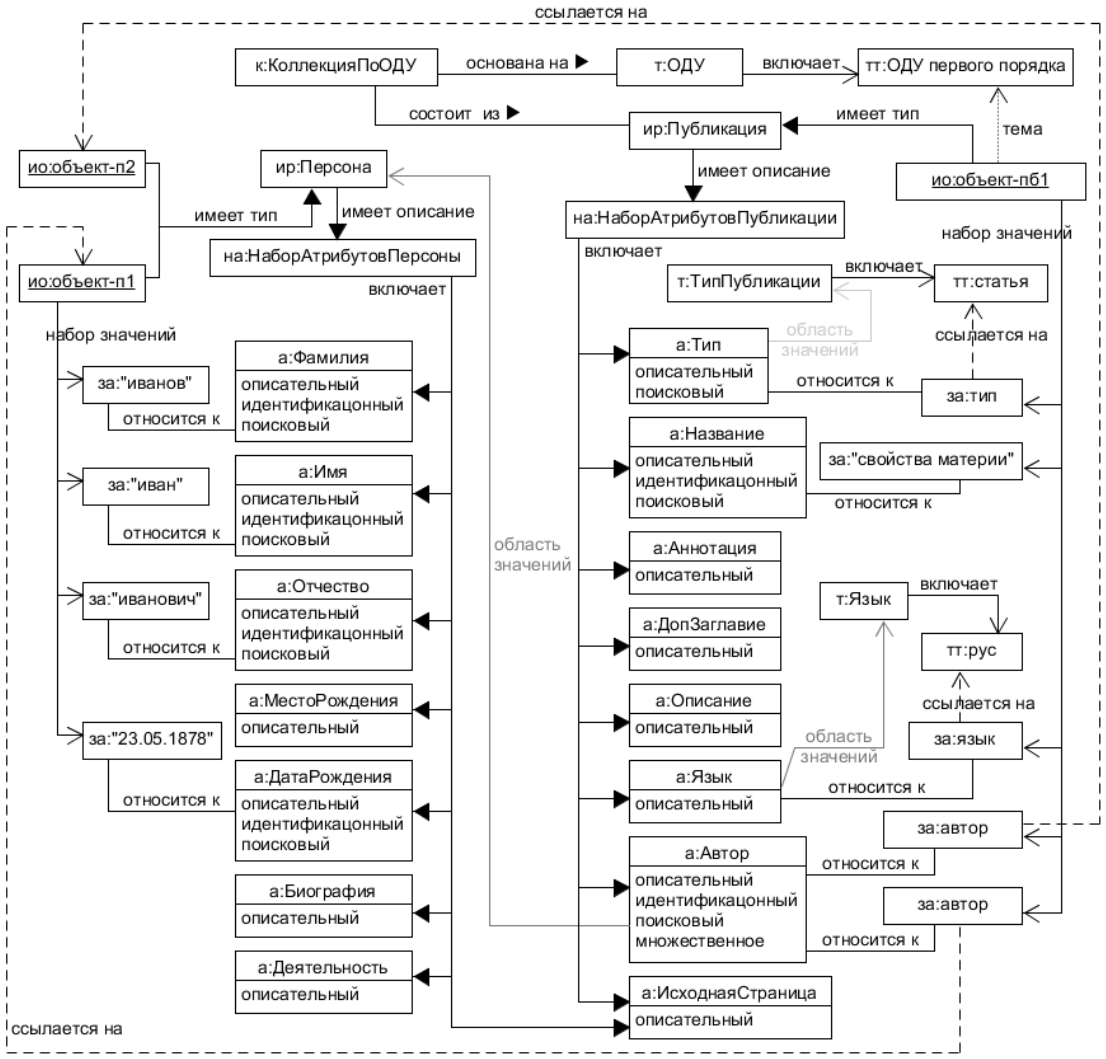

Рисунок 14

Конкретные персоны и объекты, представляющие экземпляры класса *информационный объект,* определяются информационным типом ресурса и представляются значениями свойств атрибутов соответствующего ресурса. Помимо свойств, представленных в наборе атрибутов своего типа, каждый объект обладает также свойствами, общими для всех информационных объектов, а именно, *теги*, *описание*, *дата создания, дата изменения, владелец, уникальный идентификатор.*

На рисунке 14 представлена схематически структура контента для описанного случая. На схеме проиллюстрированы связи между экземплярами информационных ресурсов *персона* и *публикация* и конкретными экземплярами класса информационного объекта (названия объектов *объект-п1, объект-п2, объект-пб1* подчеркнуты). Префиксы, отделяемые двоеточием «*ио*», «*ир*», «*к*»,

«*т»*, «*тт»*, «*на»*, «*а»,* указывают на принадлежность экземпляра соответственно к классам *информационный объект, информационный ресурс, коллекция, таксономия/тезаурус, таксон/понятие, набор атрибутов, атрибут*.

Серые стрелки, исходящие из экземпляров атрибутов, указывают на область возможных значений для него. Областью значений остальных атрибутов являются простые типы данных. На схеме значения атрибутов представлены с помощью объектов вспомогательного класса *значение атрибута* с префиксом «*за»*. Объекты этого класса содержат для простых типов атрибутов их значения (например, значения текстовых атрибутов *фамилия, имя, название* представлены на схеме в кавычках). Для объектного атрибута *автор* его значение содержит ссылку на соответствующий экземпляр информационного объекта с типом *персона*, что отображено на схеме пунктирной стрелкой. Таксономические атрибуты *тип, язык* в качестве области значений указывают на соответствующие таксономии *тип публикации* и *язык,* представляющие собой линейные словари, элементы которых (таксоны) используются в качестве значений атрибутов.

Для каждого атрибута указан его вид *описательный, идентификационный* или *поисковый*. Атрибут может относиться к нескольким видам одновременно. Поисковые атрибуты используются для динамической генерации формы поиска по объектам определенного типа ресурсов. Описательные атрибуты используются для генерации формы представления информации об объекте для пользователя. Идентификационные атрибуты, базовый набор значений которых необходим, как понятно из названия, для идентификации объекта. В наборе атрибутов для публикаций атрибут *автор* помечен как *множественный*. Этот атрибут может иметь при описании информационных объектов соответствующих по типу ресурса публикациям несколько значений, что отражено в качестве примера на схеме.

Особенности математической предметной области

Для адаптации LibMeta под математическую область знаний с учетом вышеописанных особенностей представления математических формул в

представление понятия тезауруса добавляется свойство ***formula***, которое содержит формулу на языке LaTeX, как в исходной редакции тезауруса.

В настоящее время поиск по тексту стал массовым явлением. Множество поисковых систем готовы ответить на любой вопрос пользователя. При этом с поиском математических формул дело обстоит не столь гладко. Текстовый поиск, скорее всего, не даст нужного результата, поскольку формулы в сети Интернет и в различных информационных системах представляются в виде картинок или специализированных форматов – MathML, LaTeX, Open Math [28, 29, 30, 130]. Нужна поисковая подсистема, которая учитывала бы семантику формулы, и давала пользователю возможность самому ввести формулу для поиска.

Для этого в систему была подбавлена подсистема, которая позволяет производить поиск по формулам с учётом их семантики. Для поддержки поиска по формулам в подсистеме было введено понятия *Формула* которое позволяет хранить оригинальную строку формулы из источника откуда она получена. Строка может быть в формате Content MathML, Presentation MathML, LATeX. При необходимости количество типов представления формулы в различных нотациях легко расширяется. Это понятие *Формулы* связано отношениями с *Информационными объектами* и *понятиями* тезауруса. Таким образом, мы всегда можем построить сеть связей формулы с различными информационными объектами системы. Каждая формула может быть дополнена ключевыми словами. Ключевые слова могут проставляться как экспертом системы, так и добавляться автоматически, поступая вместе с формулой из ее источника, а так же пополняясь ключевыми словами связанных объектов.

Что касается непосредственно интерфейса поиска, мы воспользовались open-source редактором математических формул VisualMathEditor, автор – David Grima [31] . Этот редактор позволяет пользователю набирать формулу, не зная тонкостей работы с LaTeX или MathML. В интерфейс добавлена кнопка поиска и поле для ввода ключевых слов.

Поиск по формулам состоит из двух логических частей – поиск собственно по формуле и поиск по ключевым словам. Искать по ключевым словам

необходимо для того, чтобы сузить круг кандидатов. Поиск по формуле должен выдавать формулы, которые полностью идентичны введённой для поиска формуле, либо содержат часть, идентичную введённой формуле. Используемый алгоритм поиска, можно условно разделить на три фазы:

1. Выбор формул - кандидатов. При необходимости преобразование формул в MathML. На этом этапе мы получаем список формул из тезауруса, которые соответствуют критериям поиска по ключевым словам.
2. Генерация внутреннего представления для формул. Для каждой формулы мы строим внутреннее представление или же используем заранее построенное внутреннее представление.
3. Сравнение искомой формулы с формулами кандидатами на полное или частичное совпадение (часть формулы-кандидата эквивалентна искомой формуле).
4. Формирование и вывод результатов поиска.

Выбор формул–кандидатов по ключевым словам происходит следующим образом: пользователь вводит ключевые слова, разделённые пробелом. В случае хотя бы одного совпадения любого ключевого слова формула попадает в список формул, подлежащих сравнению с искомой формулой.

После того, как запрос вернёт формулы-кандидаты, необходимо убедиться, что все они имеют представление в формате MathML. Если они не имеют такого представления, необходимо конвертировать формулы из LaTeX в MathML (формулы, которые не имеют ни LaTeX, ни MathML записи, не входят в результаты поиска). Для этого используется библиотека MathToWeb [32]. Для ускорения процесса будем выполняется в несколько потоков. После конвертации необходимо сохранить результаты в соответствующее поле с тем, чтобы во время следующего поиска использовать результаты конвертации.

Результаты

Средствами системы для каждой публикации на основе ее названия, аннотации и ключевых слов были выявлены связи с тезаурусом ОДУ. В качестве семантических меток были использованы термины математической

энциклопедии. Такое связывание позволило выявить с некоторой долей вероятности статьи, относящиеся к предметной области ОДУ в имеющемся набором публикаций, и организовать их в коллекции на основе тезауруса и выявленных семантических меток. Было использовано описание около 2000 публикаций, из них около 30% были отнесены в область ОДУ и имели связи со смежными областями, выявленные согласно семантическим меткам.

6.2.2.Семантическая библиотека «Задачи математической физики»

Рассмотрим в качестве следующего примера реализации семантической библиотеки на основе изложенной в работе модели, предметную область задач математической физики (далее МФ). На основе предложенной модели было выполнено конструирование библиотеки для этой области. В качестве тезауруса использован тезаурус, разработанный коллективом специалистов в этой области.

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***.

Структура тезауруса «Задачи математической физики»

Одной из основных задач при разработке тезауруса является разработка его структуры с учетом особенностей предметной области. Структура определяется тематическими разделами, наборами связей между элементами тезауруса, структурой тезаурусных статей.

Исходя из вышесказанного, базовая версия тезауруса включает в себя следующие основные тематические разделы

- *задачи математической физики,*
- *уравнения математической физики*
  - *дифференциальные уравнение с частными производными*

– *уравнения смешанного типа,*

Из анализа предметной области выявлена необходимость фиксировать следующие категории для этих разделов

- *Тип задачи (эллиптическая, гиперболическая, параболическая)*
- *Размерность задачи (одномерные, двухмерные, трехмерные)*
- *Вид уравнений (именные, нарицательные)*
- *Однородность уравнений (линейные однородные, линейные неоднородные, …)*
- *Типы коэффициентов в уравнениях (постоянные, переменные, …*)
- *Типы уравнений (эллиптическое, гиперболическое, параболическое)*

На основе этих категорий фиксируются связи

- *Вид задачи – размерность задачи*
- *Вид уравнений – тип уравнений*
- *Однородность уравнений – Типы коэффициентов в уравнениях*

Также фиксируются между терминами тезауруса:

- *иерархические: род, вид*
- *горизонтальные: синонимы, ассоциация*

Помимо основных категорий терминов в тезаурусе необходимо ввести дополнительные категории, поддерживающие возможность создания разнообразных связей с объектами, не являющимися явно понятиями тезауруса, но необходимыми для полноты описания. К таким объектам в рассматриваемом тезаурусе относятся *Авторы* и *Литература*. Для этого в структуре понятия тезауруса предусмотрен соответствующий набор связей для описания списков литературы, авторов и т.д.

- *Литература* - для возможности описания ссылок на литературу, которая содержит углубленные сведения о понятии тезауруса
- *Автор* - возможности обозначения автора понятия

Рассмотренные категории для терминов связанные выделенными в процессе анализа предметной области иерархическими и горизонтальными связями образуют концептуальную модель рассматриваемой предметной области.

В концептуальной модели тезауруса, таким образом, фиксируется:

- способ определения понятий;
- способ определения синонимов понятий;
- перечень их свойств и атрибутов понятия;
- категории объектов;
- состав объектов каждой категории.

Таким образом, структурно понятие тезауруса включает в себя следующие элементы:

- буквенно-цифровой код понятия;
- дескриптор понятия;
- недескрипторы – синонимы понятия;
- тематический раздел понятия;
- символьное представление формулы понятия;
- перечень связей с другими понятиями;
- текстовые дополнения (комментарии, замечания, справки);
- список литературы для понятия;
- авторы понятия.

Из сущности описания структуры тезауруса появляется необходимость фиксировать также основные связи различных лексико-семантических категорий, таких как:

- вид уравнения: одномерное, двухмерное, трехмерное;
- тип уравнения: гиперболическое, параболическое, эллиптическое;
- типы коэффициентов: переменные, постоянные;
- и т.д.

Онтология тезауруса «Задачи математической физики»

Описание тезауруса «уравнений математической физики» в терминах понятий базовой версии онтологии расширяется дополнительно с помощью понятий расширенной модели для возможности расширения структуры статьи понятия этого тезауруса. Добавляются такие атрибуты как: *комментарий, замечание, справка, литература, авторы*.

Рассмотренные выше атрибуты *замечание, справка* представляют собой экземпляры классов ***ThesaurusAttributeText,*** *комментарий,* - ***ThesaurusAttributeString***, а *литература, авторы* - представляют собой экземпляры классов ***ThesaurusAttributeObject***. В свою очередь они объединены в набор атрибутов для данного тезауруса.

Таким образом, мы переходим к трехуровневому представлению тезауруса предметной области в рамках библиотеки LibMeta. Для построения тезауруса выполняется последовательность шагов 1 – 4 в рамках предложенного подхода, описанных в разделе 5.1.

Онтология контента ресурсов «Задачи математической физики»

Процесс формирования тезауруса основан на анализе первоисточников классиков математического анализа и дифференциального исчисления, для чего был сформирован представительный список публикаций. Соответственно в качестве основного контента библиотеки рассматривался этот список. В качестве конструируемых типов ресурсов для первоначальной версии, рассматривались *Авторы* и *Публикации*. Был определен набор атрибутов для каждого типа ресурсов в рамках минимального набора свойств на основе Dublin Core для публикаций и FOAF для описания авторов. Во многом процесс моделирования контента совпадает с аналогичным процессом выполненным в предыдущем примере для области ОДУ.

Особенности математической предметной области

Так же как и в предыдущем примере, область знаний математической физики не представима без использования формул. С учетом вышеописанных

особенностей представления математических формул в представлении понятия тезауруса добавлялись свойства, которые содержит формулу на языке LaTeX, как в исходной редакции тезауруса. Для поддержки поиска по формулам был реализован дополнительный модуль поиска учитывающий специфику, как их синтаксиса, так и набора связей формул, а также как и для предметной области ОДУ, был введен отдельный тип информационных ресурсов *«Формула»* для отображения всех возможных ее семантических связей.

Результаты

Средствами системы каждой публикации были сопоставлены связи с понятиями тезауруса МФ. В качестве семантических меток так же были использованы термины математической энциклопедии. Было использовано описание около 3000 публикаций, из них около 40% были отнесены в область МФ и имели связи со смежными предметными областями, выявленные согласно семантическим меткам.

6.2.3.Семантическая библиотека «Микробиология и физиология растений»

В качестве следующего примера реализации семантической библиотеки, на основе изложенной в работе модели, рассмотрим предметную область «Микробиология и физиология растений» (далее МиФР). На основе предложенной модели было выполнено конструирование библиотеки для этой области. В качестве тезауруса использован тезаурус, разработанный коллективом специалистов в этой области.

В качестве источника данных для поступающих в систему публикации выступили реляционные БД, находящиеся в библиотеке БЕН РАН.

Структура тезауруса «Микробиология и физиология растений»

Структура рассматриваемого тезауруса 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. Помимо основных понятий в тезаурусе необходимо ввести дополнительные категории, поддерживающие возможность сохранения информации об источниках данных, откуда эти понятия извлекались. Сами источники не являются явно понятиями тезауруса, но необходимыми для полноты описания. Для этого в тезаурусе вводятся связи с ресурсом *Источник Понятия*. Для этого в структуре понятия тезауруса предусмотрен соответствующий атрибут *источник*.

Онтология тезауруса «Микробиология и физиология растений»

Описание тезауруса «*Микробиология и физиология растений*» в терминах понятий базовой версии онтологии расширяется дополнительно с помощью понятий расширенной модели для возможности расширения структуры статьи понятия этого тезауруса. Добавляются такие атрибуты как: *определение, код УДК, код ББК, источник (в виде ссылки), источник (в виде строки), смежная предметная область*.

Рассмотренные выше атрибуты *определение* представляют собой экземпляры классов ***ThesaurusAttributeText,*** *смежная предметная область , источник (в виде строки)* – ***ThesaurusAttributeString***, *источник (в виде ссылки)* – ***ThesaurusAttributeHref***, а *код УДК, код ББК* – представляют собой экземпляры классов ***ThesaurusAttributeTaxonomy***. В свою очередь они объединены в набор атрибутов для данного тезауруса.

Также между понятиями устанавливаются дополнительные горизонтальные связи *«содержит в», «входит в», «эквивалентно»* - как экземпляры класса ***HierarchicalRelation***.

Онтология контента ресурсов «Микробиология и физиология растений»

В качестве контента системы используются научные публикации по микробиологии. Их структура была смоделирована с помощью классов *InformationResourse, Attribute, AttributeSet* как и в случае с примером моделирования контента в предметной области ОДУ.

Результаты

Средствами системы каждой публикации были сопоставлены связи с понятиями тезауруса МиФР. Были использованы описания около 3000 публикаций, из области МиФР. Для них были проставлены связи с терминами тезауруса и выявлены связи со смежными предметными областями, выявленные согласно дополнительным сведениям из тезауруса.

6.2.4.Семантическая библиотека «Математическая энциклопедия»

В данной реализации семантической библиотеки строится электронная версия советской математической энциклопедии 1978 года. Это справочное издание по всем разделам математики, основу которого составляют статьи, посвященные важнейшим направлениям математики. Принцип расположения статей в энциклопедии — алфавитный. В ней широко используется система ссылок на другие статьи. В отличие от ее использования в первых двух примерах, где использовались только ее термины из предметного указателя без ссылок на статьи, в данной реализации было рассмотрено все богатство связей между ее понятиями. Также для ее расширения использовалась англоязычная версия этой энциклопедии, а именно использовались коды *MSC* проставленные для понятий в англоязычной версии и текстовое представление формул в виде *TEX* нотаций.

В данной задаче в качестве внешнего источника данных была использована система DBPedia, для выявления дополнительных связей.

Структура тезауруса «математическая энциклопедия»

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.

Онтология тезауруса «математическая энциклопедия»

Для описания онтологии тезауруса математической энциклопедии базовая версия тезауруса была расширена дополнительно с помощью понятий расширенной модели. Добавляются такие атрибуты как: *формула, персона, код УДК, код MSC, ссылка (на англоязычную версию понятия)*.

Рассмотренные выше атрибуты *ссылка* представляют собой экземпляры классов ***ThesaurusAttributeHref,*** *формула, персона* **- *ThesaurusAttributeObject***. В свою очередь они объединены в набор атрибутов для данного тезауруса, *код УДК, код MSC* – экземпляры классов ***ThesaurusAttributeTaxonomy.***

Онтология контента ресурсов «математическая энциклопедия»

В качестве контента библиотеки использовались массивы научных математических публикаций накопленных за время работы в первых двух системах с соответствующей структурой, а так же использовался тип информационных ресурсов *Формула*, по аналогии с примером из предметной области ОДУ.

Результаты

Средствами системы для каждой публикации на основе ее названия, аннотации и ключевых слов были выявлены связи с терминами математической энциклопедии. Это позволило дополнительно произвести дополнительное тематическое разбиение публикаций в рамках предметной области. Такое связывание позволило выявить с некоторой долей вероятности статьи,

относящиеся к разным разделам математики, и организовать их в коллекции на основе тезауруса и проставленных связей MSC. Было использовано около 5000 публикаций.

6.2.5.Подключение реляционных источников

При реализации перечисленных примеров семантических библиотек на основе LibMeta практически во всех случаях вставал вопрос подключения в качестве источников данных реляционные базы данных [41, 42, 43]. 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. Детали подключения выходят за рамки этой работы, с документацией можно ознакомиться по ссылкам [33, 34].

6.3. RESTfull API

Рассмотрим организацию 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. Чаще всего используется для получения данных и интеграции с внешними системами. 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 [35, 36]. Удобством использования REST является возможность проверки получаемых данных в

обычном браузере, либо использованием стандартных приложений для выполнения HTTP запросов [37, 38, 40].

Выбор операций, которые должны поддерживаться прикладным интерфейсом, не всегда является простой задачей, так как только интерфейс будет опубликован, следует избегать изменений в нем, чтобы не нарушить работоспособность использующих его клиентов. Добавление новых операций в прикладной интерфейс обычно вызывает меньше проблем, чем изменение или удаление существующих. Поэтому сначала был реализован минимально необходимый набор операций, а затем было принято решение развивать интерфейс по мере необходимости.

Клиенты должны иметь следующие возможности:

- Получить список существующих информационных ресурсов;
- Создавать новые информационные объекты;
- Изменить существующие информационные объекты;
- Извлекать список существующих информационных объектов, соответствующих определенному типу ресурсов;
- Извлекать описание информационного ресурса;
- Извлекать содержимое информационного объекта;
- Извлекать описание тезауруса;
- Извлекать список понятий тезауруса;
- Создавать новые понятия тезауруса
- Изменить существующие понятия тезауруса
- Извлекать содержимое понятия тезауруса;

Браузер взаимодействует с нашим приложением с помощью платформонезависимого протокола HTTP на основе которого реализован архитектурный стиль REST который может использоваться в любых взаимодействиях типа клиент/сервер. Основная идея заключается в том, что приложение рассматривается как коллекция ресурсов, над которыми можно выполнять некоторые операции с помощью нескольких методов.

Возьмем в качестве примера информационный объект, который является отдельным элементом контента библиотеки, который можно извлечь, изменить или удалить. На рисунке 15 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 а RDF/XML – разметке, возвращаемого клиенту.

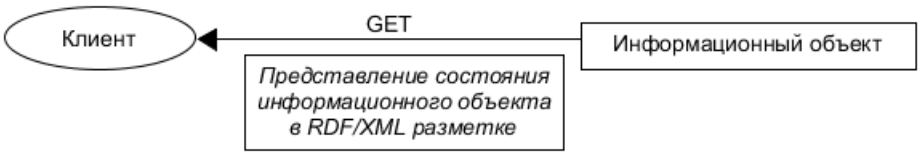

Рисунок 15

Как видно главное преимуществ 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).

Четырьмя основными операциями (или методами), поддерживаемыми протоколом HTTP, являются GET, POST, PUT и DELETE:

- Метод GET запрашивает представление информационного *ресурса/информационного объекта/тезауруса/понятия тезауруса*. Запросы с использованием этого метода могут только извлекать данные.
- POST используется для отправки сущностей определённого типа (*ресурса/информационного объекта/тезауруса/понятия тезауруса*). Часто вызывает изменение состояния или какие-то другие эффекты на сервере.
- PUT заменяет все текущие представления *информационного объекта/понятия тезауруса* данными запроса.
- DELETE используется для удаления *информационного объекта*.

Часто при реализации технологии REST возникает недопонимание, когда следует применять метод POST, а когда PUT. Согласно спецификации метод PUT используется для создания и обновления ресурсов с заданными адресами URL. В большинстве случаев в нашей системе идентификатор нового ресурса неизвестен до его создания, следовательно, определить конкретный адрес URL для метода

PUT невозможно. По этой причине при выполнении метода POST, будет создаваться объект, если URL объекта в системе не существует. Если же URL объекта существует, то будет выполняться изменение этого объекта.

Чтобы получить доступ к ресурсу, необходимо указать его адрес URL. Адреса должны быть:

- *Уникальными* для каждого ресурса – каждому адресу URL должен соответствовать единственный ресурс.
- *Долгоживущими* – адрес URL всегда должен указывать на один и тот же ресурс. Это жесткое требование, и спецификация HTTP признает это, предусматривая временное или постоянное перенаправление.

Рассмотрим примеры запросов для информационных ресурсов, информационных объектов, тезауруса и понятий тезауруса.

Точкой входа для доступа к RESTfull API в нашей системе является URL вида:

```
http(s)://[имя сервера]/rest/
```

Все методы объединены в группы:

`/thesaurus` – тезаурус

`/resource` – информационные ресурсы

`/objects` – информационные объекты

`/object` – информационный объект

`/concept` – понятия тезауруса

Информационные ресурсы

- Получить список существующих информационных ресурсов

Для извлечения списка используемых информационных ресурсов необходимо сформировать GET запрос по адресу:

```
http(s)://[имя сервера]/rest/resource/all
```

- *Извлекать описание информационного ресурса*

Для извлечения описания информационного ресурса необходимо сформировать GET запрос по адресу:

```
http(s)://[имя сервера]/rest/resource/<id>
```

где <id> идентификатор запрашиваемого ресурса

Информационные объекты

- *Извлекать список информационных объектов, соответствующих определенному типу ресурсов*

  Для извлечения списка информационных объектов соответствующих определенному типу ресурсов необходимо сформировать GET запрос по адресу:

```
http(s)://[имя сервера]/rest/objects/<rid>
```

  где <rid> идентификатор информационного ресурса запрашиваемых объектов

- *Извлекать содержимое информационного объекта*

  Для извлечения содержимого информационного объекта необходимо сформировать GET запрос по адресу:

```
http(s)://[имя сервера]/rest/object/<id>
```

  где <id> идентификатор запрашиваемого информационного объекта

- *Создание новых информационных объектов*

  Для создания нового информационного объекта необходимо сформулировать POST запрос вида

```
http(s)://[имя сервера]/rest/object/
```

- *Изменнение существующие информационные объекты;*

  Для изменения существующего информационного объекта необходимо сформулировать POST запрос вида:

```
http(s)://[имя сервера]/rest/object/<id>
```

где <id> идентификатор изменяемого информационного объекта

Тезаурус

- *Извлекать описание тезауруса*

  Для извлечения описания тезауруса необходимо сформировать GET запрос по адресу:

```
http(s)://[имя сервера]/rest/thesaurus
```

Понятия тезауруса

- *Извлекать список понятий тезауруса*

  Для извлечения списка понятий тезауруса необходимо сформировать GET запрос по адресу:

```
http(s)://[имя сервера]/rest/concept/all
```

- *Извлекать содержимое понятия тезауруса*

  Для извлечения содержимого понятия тезауруса необходимо сформировать GET запрос по адресу:

```
http(s)://[имя сервера]/rest/concept/<cid>
```

где <cid> идентификатор запрашиваемого понятия тезауруса

- *Создавать новые понятия тезауруса*

  Для создания нового понятия тезауруса необходимо сформулировать POST запрос по адресу:

```
http(s)://[имя сервера]/rest/concept
```

- *Изменение существующие понятия тезауруса*

  Для изменения существующего понятия тезауруса необходимо сформулировать POST запрос по адресу:

```
http(s)://[имя сервера]/rest/ concept/<cid>
```

где <cid> идентификатор изменяемого понятия тезауруса

В случае, если указан неверный адрес - возвращается ошибка 404. В случае возникновения ошибки авторизации, возвращается ошибка 401. Функции, возвращающие массив значений, поддерживают параметры для порционной загрузки, которые передаются в виде query-строки(?top=1&skip=10)

Для работы с функциями загрузки и обновления данных через REST API пользователю необходимо пройти авторизацию и пользователь должен обладать правами «REST подключение» или «Администратор». В системе реализована BASIC авторизация, которая используется браузером для доступа к функциям API. При использовании данного метода необходимо в заголовке каждого запроса указывать: Authorization: Basic {login}:{password}. Допускается использование стандарта кодирования двоичных данных base64 при формировании строки {login}:{password}

6.4. Дальнейшее развитие

Дальнейшее развитие системы LibMeta включает в себя

- реализацию функции формирования рефератов по различным направлениям предметной области на основе ее тезауруса.
- решается задача расширения подсистемы аннотирования текстов для специализированных предметных областей, например, математических, улучшив возможности семантической обработки формул

- работа с междисциплинарными областями знаний и формирования семантических связей между разными предметными областями и навигацией по связанным ресурсам
- разработка модуля автоматизированного расширения тезауруса конкретной предметной области.

6.5. Выводы

В настоящей главе представлено описание работы прототипа системы LibMeta, которая разработана автором работы на основе предложенного в диссертации подхода к разработке семантических библиотек.

7. Заключение

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.

Предложенный 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.

Предложен способ интеграции данных библиотеки с внешними источниками удовлетворяющих требованиям LOD, являющемуся практическим воплощением парадигмы Semantic Web.

Литература

1. Palano R., Pandurino A., Guido A. L. Conceptual design of web application families: the BWW approach //Proceedings of 6th Workshop on Domain Specific Modeling, Portland, USA. – 2006. – С. 23-32.
2. Candela L., Castelli D., Dobreva M., Ferro N., Ioannidis Y., Katifori H., Koutrika G., Meghini C., Pagano P., Ross S., Agosti M., Schuldt H., Soergel D. The DELOS Digital Library Reference Model Foundations for Digital Libraries. IST–2002 2.3.1.12. Technology–enhanced Learning and Access to Cultural Heritage. Version 0.98, December 2007. http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf
3. Kruk S. R. et al. JeromeDL-a Semantic Digital Library. – 2007.
4. Miller G. A. WordNet: a lexical database for English //Communications of the ACM. – 1995. – Т. 38. – №. 11. – С. 39-41.
5. Kruk S. R., Synak M., Zimmermann K. MarcOnt--Integration ontology for bibliographic description formats //International Conference on Dublin Core and Metadata Applications. – 2005. – С. pp. 231-234.
6. FOAF Vocabulary Specification 0.99 // http://xmlns.com/foaf/spec/.
7. Isaac A., Summers E. Skos simple knowledge organization system primer //Working Group Note, W3C. – 2009.
8. SPARQL 1.1 Overview // https://www.w3.org/TR/sparql11-overview/
9. http://www.europeana.eu.
10. Doerr M. et al. The europeana data model (edm) //World Library and Information Congress: 76th IFLA general conference and assembly. – 2010. – С. 10-15.
11. Lehmann J. et al. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia //Semantic Web. – 2015. – Т. 6. – №. 2. – С. 167-195..
12. Серебряков В. А., Что такое семантическая цифровая библиотека // Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL. - Дубна, Объединенный институт ядерных исследований. - 2014. - С. 21- 25.
13. Грегер С. Э., Поршнев С. В. Построение онтологии архитектуры информационной системы //Фундаментальные исследования. – 2013. – Т. 11. – №. 10.
14. ГОСТ 7.0-99 Межгосударственный стандарт ГОСТ 7.0-99 "Система стандартов по информации, библиотечному и издательскому делу. Информационно-библиотечная деятельность, библиография. Термины и определения" (введен в действие постановлением Госстандарта РФ от 7 октября 1999 г. N 334-ст)
15. Губанов Николай Иванович, Губанов Николай Николаевич, Волков Андрей Эдуардович. "Критерии истинности и научности знания" Философия и общество, no. 3 (80), 2016, pp. 78-95.: https://cyberleninka.ru/article/n/kriterii-istinnosti-i-nauchnosti-znaniya
16. Ильин В. В., Калинкин А. Т. Природа науки: Гносеологический анализ. М.: Высшая школа, 1985. – 230 с.
17. Semantic Web // http://www.w3.org/standards/semanticweb/
18. Kumazawa Terukazu. "Toward knowledge structuring of sustainability science based on ontology engineering". Sustainability Science. 4: 99–116. doi:10.1007/s11625-008-0063-z. Retrieved 22 April 2015.
19. Mascardi V., Cordi V., P Rosso (2007).A Comparison of Upper Ontologies
20. Leon Welick, Joseph W. Yode, Rebecca Wirfs-Broc Adaptive Object-Model Builder – AdaptiveObjectModel.com, 2009 – http://joeyoder.com/PDFs/04welicki.pdf

21. Joseph W. Yoder, Federico Balaguer, Ralph Johnson Architecture and Design of Adaptive Object-Model – AdaptiveObjectModel.com, 2000, http://www.adaptiveobjectmodel.com/OOPSLA2001/AOMIntriguingTechPaper.pdf
22. Чернышов В. Н., Чернышов А. В. Теория систем и системный анализ //Тамбов: Изд-во Тамб. гос. техн. ун-та. – 2008.
23. Волкова В. Н., Денисов А. А. Теория систем и системный анализ //М.: Юрайт. – 2010.
24. Т.Т. Газизов Моделирование систем // http://koi.tspu.ru/koi_books/gazizov/index.htm
25. Zhao L., Ichise R. Ontology integration for linked data //Journal on Data Semantics. – 2014. – Т. 3. – №. 4. – С. 237-254.
26. Zhao, Lihua & Ichise, Ryutaro. Integrating Heterogeneous Ontology Schema from LOD //The 26th Annual Conference of the Japanese Society for Arti_cial Intelligence – 2012.
27. Владикавказский математический журнал. http://www.vmj.ru.
28. OpenMath and MathML: Semantic Mark Up for Mathematics. [Электронный ресурс]. - Электрон. дан. – URL: http://www.acm.org/crossroads/xrds6-2/openmath.html
29. Елизаров А. М., Липачѐв Е. К., Малахальцев М. А. Основы MathML Представление математических текстов в Internet. – 2008.
30. LaTeX – A document preparation system https://www.latex-project.org/
31. Visual Math Editor // http://visualmatheditor.equatheque.net/
32. Math To Web – User's Guide. [Электронный ресурс]. - Электрон. дан. – URL: http://www.mathtoweb.com/cgi-bin/mathtoweb_users_guide.pl
33. Bizer C., Cyganiak R. D2r server-publishing relational databases on the semantic web //Poster at the 5th international semantic web conference. – 2006. – Т. 175.
34. Accessing Relational Databases as Virtual RDF Graphs // http://d2rq.org/
35. Fielding R. T., Taylor R. N. Architectural styles and the design of network-based software architectures. – Doctoral dissertation : University of California, Irvine, 2000. – Т. 7.
36. Pérez S. et al. RESTful, resource-oriented architectures: a model-driven approach //International Conference on Web Information Systems Engineering. – Springer, Berlin, Heidelberg, 2010. – С. 282-294.
37. Fensel D. et al. Web2. 0 and RESTful Services //Semantic Web Services. – Springer, Berlin, Heidelberg, 2011. – С. 67-86.
38. Hernández A. G., García M. N. M. Metadata Architecture in RESTful Design //REST: From Research to Practice. – Springer, New York, NY, 2011. – С. 459-471.
39. Berners-Lee T., Fielding R., Masinter L. Uniform resource identifier (URI): Generic syntax. – 2004. – №. RFC 3986.
40. Wilde E., Pautasso C. (ed.). REST: from research to practice. – Springer Science & Business Media, 2011.
41. Sahoo S. S. et al. A survey of current approaches for mapping of relational databases to RDF //W3C RDB2RDF Incubator Group Report. – 2009. – Т. 1. – С. 113-130
42. Michel F. et al. xR2RML: Relational and non-relational databases to RDF mapping language : дис. – CNRS, 2017.
43. Bytyçi E., Ahmedi L., Gashi G. RDF Mapper: Easy Conversion of Relational Databases to RDF //WEBIST. – 2018. – С. 161-165.
44. Horrocks I. et al. SWRL: A semantic web rule language combining OWL and RuleML //W3C Member submission. – 2004. – Т. 21. – №. 79. – С. 1-31.
45. Heath T., Bizer C. Linked data: Evolving the web into a global data space //Synthesis lectures on the semantic web: theory and technology. – 2011. – Т. 1. – №. 1. – С. 1-136.
46. Bizer C., Heath T., Berners-Lee T. Linked data: The story so far //Semantic services, interoperability and web applications: emerging concepts. – IGI Global, 2011. – С. 205-227.
47. Schmachtenberg M., Bizer C., Paulheim H. State of the LOD Cloud 2014 //University of Mannheim, Data and Web Science Group. August. – 2014. – Т. 30.

48. Костин В. В. Обзор семантических моделей, описывающих научные публикации и научно-исследовательскую деятельность // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. 2014.
49. Miles A. et al. SKOS core: simple knowledge organisation for the web //International Conference on Dublin Core and Metadata Applications. – 2005. – С. 3-10.
50. Ле Х., Тузовский А. Ф. Разработка семантических электронных библиотек //Доклады Томского государственного университета систем управления и радиоэлектроники. – 2011. – №. 2-2 (24).
51. Когаловский М. Р., Паринов С. И. Семантическое структурирование контента научных электронных библиотек на основе онтологий. – 2015.
52. Яковлева М. В., Тен А. К., Куглер В. М. На пути к созданию электронной семантической библиотеки //Электронные библиотеки: перспективные методы и технологии, электронные коллекции». XIII Всероссийская научная конференция RCDL. – 2011. – С. 400-401.
53. Sadeh T., Walker J. Library portals: toward the semantic Web //New Library World. – 2003. – Т. 104. – №. 1/2. – С. 11-19.
54. DCMI Home: Dublin Core® Metadata Initiative (DCMI). http://dublincore.org/
55. Crofts N., Doerr M., Gill T., Stead S., Stiff M. (editors), Definition of the CIDOC Conceptual Reference Model, January 2008. Version 4.2.4.
56. Hammond, T. (2008). RDF Site Summar. http://www.w3.org/TR/owl2-overview/y 1.0 Modules: PRISM. http://nurture.nature.com/rss/modules/mod_prism.html
57. International Digital Enterprise Alliance (2009). Publishing Requirements for Industry Standard Metadata Specification Version 2.0. Alexandria, VA, USA: IDEAlliance.
58. Bibliographic Ontology. http://bibliontology.com/
59. Brickley D., Miller L. FOAF vocabulary specification 0.91. – 2007.
60. Graves M., Constabaris A., Brickley D. Foaf: Connecting people on the semantic web //Cataloging & classification quarterly. – 2007. – Т. 43. – №. 3-4. – С. 191-202.
61. Лаврёнова О. А., Павлов В. В. Библиотечно-библиографическая классификация как традиционная система организации знаний в среде открытых связанных данных //Научные и технические библиотеки. – 2017. – №. 4. – С. 44-60.
62. Шварцман М. Е., Найдин О. П. Linked Open Data как средство обогащения поисковых запросов //Унив. кн. – 2015. – №. 12. – С. 66-71.
63. Guarino N. Formal Ontology and Information Systems // N. Guarino (ed.) Formal Ontology and Information Systems. – Amsterdam, 1998. – С. 3–15.
64. Hruby P. Ontology-Based Domain-Driven Design // [Soft-MetaWare] URL: www.softmetaware.com/oopsla2005/hruby.pdf.
65. Леонова Ю. В., Федотов А. М. Создание прототипа системы управления информационными ресурсами //Вестник Восточно-Казахстанского гос. Техн. Университета и журнала Вычислительные технологии ИВТ СО РАН.–CITech-2018, Усть-Каменогорск, Казахстан. – 2018. – С. 47-56.
66. Кулагин М. В., Лопатенко А. С. Научные информационные системы и электронные библиотеки. Потребность в интеграции //Электронные библиотеки: перспективные методы и технологии, электронные коллекции. – 2001.
67. Федотов А. М., Шокин Ю. И. Электронная библиотека Сибирского отделения РАН // Информационное общество. — 2000. — № 2. — С.22–31.
68. Шокин Ю. И., Федотов А. М., Жижимов О. Л., Федотова О. А. Эволюция информационных систем: от Web-сайтов до систем управления информационными ресурсами // Вестник НГУ Серия: Информационные технологии. 2015. Том 13, Выпуск № 1. С. 117–134. — ISSN 1818-7900.
69. Каленов Н. и др. Электронная библиотека Научное наследие России //Бібліотечний вісник. – 2009. – №. 6. – С. 40-42.

70. Вигурский К., Горный Е. Развитие электронных библиотек: мировой и российский опыт, проблемы, перспективы. – 2015.
71. Когаловский М. Р. Метаданные, их свойства, функции, классификация и средства представления. – Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции»—RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г., 2012.
72. Hlava M. M. K. The Taxobook: History, Theories, and Concepts of Knowledge Organization, Part 1 of a 3-Part Series //Synthesis Lectures on Information Concepts, Retrieval, and Services. – 2014. – Т. 6. – №. 3. – С. 1-80.
73. Hlava M. M. K. The taxobook: Principles and practices of building taxonomies, part 2 of a 3-part series //Synthesis Lectures on Information Concepts, Retrieval, and Services. – 2014. – Т. 6. – №. 4. – С. 1-164.
74. Hlava M. M. K. The Taxobook: Applications, Implementation, and Integration in Search: Part 3 of a 3-Part Series //Synthesis Lectures on Information Concepts, Retrieval, and Services. – 2014. – Т. 6. – №. 4. – С. 1-156.
75. Cardillo E. et al. Towards the Reuse of Standardized Thesauri Into Ontologies //WOP. – 2014. – С. 26-37.
76. Лукашевич Н. В. Тезаурусы в задачах информационного поиска / Москва: Издательство МГУ. — 2011. — С. 512. — ISBN 978-5-211-05926-9.
77. Жижимов О. Л., Мазов Н. А., Федотов А. М. Некоторые заметки об эволюции цифровых репозитариев традиционных библиотек к полнофункциональным электронным библиотекам //Территория новых возможностей. Вестник Владивостокского государственного университета экономики и сервиса. – 2010. – №. 3 (7).
78. Noruzi A. Folksonomies:(un) controlled vocabulary? //KO KNOWLEDGE ORGANIZATION. – 2006. – Т. 33. – №. 4. – С. 199-203.
79. Vander Wal T. Folksonomy. – 2007.
80. Gruber T. Ontology of folksonomy: A mash-up of apples and oranges //International Journal on Semantic Web and Information Systems (IJSWIS). – 2007. – Т. 3. – №. 1. – С. 1-11.
81. Kless D. et al. Thesaurus and ontology structure: Formal and pragmatic differences and similarities //Journal of the Association for information science and technology. – 2015. – Т. 66. – №. 7. – С. 1348-1366.
82. Гуревич И.Б. , Трусова Ю.О. Тезаурус и онтология предметной области “Анализ изображений” // Всероссийская конф. с междунар. участием "Знания – Онтологии – Теории" (ЗОНТ–09). – Новосибирск: Институт математики им. С.Л. Соболева СО РАН, 2009. – 10 с.
83. Кузнецова А. Г. Программные комплексы для электронных библиотек //Молодежный вестник Санкт-Петербургского государственного института культуры. – 2015. – №. 1. – С. 43-47.
84. Don K. J., Bainbridge D., Witten I. H. The design of Greenstone 3: An agent based dynamic digital library. – Technical report, Department of Computer Science, University of Waikato, Hamilton New Zealand, 2002.
85. Don K. Greenstone3: A modular digital library //User and developer manual. Work in progress.(Last updated 9/11/05) Online: http://www. greenstone. org/greenstone3. html. – 2006.
86. Witten I. H., Bainbridge D. Creating digital library collections with Greenstone //Library hi tech. – 2005. – Т. 23. – №. 4. – С. 541-560.
87. Барахнин В. Б., Нехаева В. А., Федотов А. М. О задании меры сходства для кластеризации текстовых документов //Вестник Новосибирского государственного университета. Серия: Информационные технологии. – 2008. – Т. 6. – №. 1.

88. Alexiev V., Isaac A., Lindenthal J. On the composition of ISO 25964 hierarchical relations (BTG, BTP, BTI) //International Journal on Digital Libraries. – 2016. – Т. 17. – №. 1. – С. 39-48.
89. ISO 25964 thesaurus schemas. http://www.niso.org/schemas/iso25964
90. De Smedt, J., Isaac, A., Clarke, S.D., Lindenthal, J., Zeng, M.L., Tudhope, D.S.,Will, L., Alexiev, V.: ISO 25964 part 1: thesauri for information retrieval: RDF/OWL vocabulary, extension of SKOSand SKOS-XL (2013). http://purl.org/iso25964/skos-thes
91. Dextre Clarke S. G., Zeng M. L. From ISO 2788 to ISO 25964: The evolution of thesaurus standards towards interoperability and data modelling //Information Standards Quarterly (ISQ). – 2012. – Т. 24. – №. 1.
92. ISO 25964-2:2013 Information and documentation—thesauri and interoperability with other vocabularies—part 2: interoperability with other vocabularies (2013). http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53658
93. Митрофанова О.А., Константинова Н.С. Онтологии как системы хранения знаний // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационнотелекоммуникационные системы». – 2008. – С. 54
94. Gruber T. R. The role of common ontology in achieving sharable, reusable knowledge bases. In J. A. Allen, R.Fikes, and E. Sandewell, editors, Principles of Knowledge Representation and Reasoning – Proceedings of the Second International Conference, pp. 601-602. Morgan Kaufmann (1991)
95. Allemang D., Hendler J. Semantic web for the working ontologist: effective modeling in RDFS and OWL. – Elsevier, 2011.
96. Богданова И. Ф., Богданова Н. Ф. Электронные библиотеки: история и современность //Труды объединённой научной конференции" Интернет и современное общество". – 2017. – №. 1. – С. 133-154.
97. Степин В. С., Горохов В. Г., Розов М. А. Философия науки и техники. – М. : Фирма" Гардарика", 1996.
98. Кошовец О. Б., Фролов И. Э. Онтология и реальность: проблемы их соотношения в методологии экономической науки //Теоретическая экономика: онтологии и этика. М.: Институт экономики РАН. – 2013. – С. 27-112.
99. Селиванов Вячеслав Михайлович Онтологические основания постмодернизма // Вестник ОГУ. 2011. №5 (124). URL: https://cyberleninka.ru/article/n/ontologicheskie-osnovaniya-postmodernizma (дата обращения: 03.06.2019).
100. Добров Б. В. И др. Онтологии и тезаурусы. – 2015.
101. В.Л. Обухов , Ю.Н. Солонин , В.П. Сальников и В.В. Василькова. Философия и методология познания: Учебник для магистров и аспирантов — Санкт-Петербургский университет МВД России; Академия права, экономики и безопасности жизнедеятельности; спбгу; спбгау; ипип (спб.) — спб.: Фонд поддержки науки и образования в области правоохранительной деятельности «Университет». — 560 с.. 2003
102. ISO 2788:1986, Guidelines for the establishment and development of monolingual thesauri, International Organization for Standardization // 1986.
103. Wand Y. Weber R., “Research Commentary: Information Systems and Conceptual Modeling – A Research Agenda,” Information Systems Research, Vol. 13, No. 4, 2002. pp. 363–376.
104. Ломов П.А., Шишаев М.Г. Интеграция онтологий с использованием тезауруса для осуществления семантического поиска // Информационные технологии и вычислительные системы. - 2009. - № 3. -С. 49-59.
105. Бездушный А.Н., Жижченко А.Б., Кулагин М.В., Серебряков В.А. Интегрированная система информационных ресурсов РАН и технология разработки цифровых библиотек. Программирование, 2000, 4, с. 3-14.

106. Katsis Y., Papakonstantinou Y. View-based data integration //Encyclopedia of Database Systems. – 2009. – С. 3332-3339.
107. Xu L., Embley D. W. Combining the Best of Global-as-View and Local-as-View for Data Integration //ISTA. – 2004. – Т. 48. – С. 123-36.
108. Когаловский М. Р. Методы интеграции данных в информационных системах //Институт проблем рынка РАН. – 2010. – Т. 74.
109. Карабач А. Е. Системы интеграции информации на основе семантических технологий //Наука, техника и образование. – 2014. – №. 2 (2).
110. Lenzerini M. Data integration: A theoretical perspective //Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. – ACM, 2002. – С. 233-246.
111. Calvanese D., De Giacomo G., Lenzerini M. Ontology of Integration and Integration of Ontologies //Description Logics. – 2001. – Т. 49. – №. 10-19. – С. 30.
112. Noy N. F. Semantic integration: a survey of ontology-based approaches //ACM Sigmod Record. – 2004. – Т. 33. – №. 4. – С. 65-70.
113. Zhao L., Ichise R. Ontology integration for linked data //Journal on Data Semantics. – 2014. – Т. 3. – №. 4. – С. 237-254.
114. Ахлёстин А.Ю., Лаврентьев Н.А., Фазлиев А.З. Систематизация научных графических ресурсов по молекулярной спектроскопии // Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (18-23 сентября 2017 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2017. — С. 34-42. — URL: http://keldysh.ru/abrau/2017/39.pdf doi:10.20948/abrau-2017-39
115. Нгок Н. Б., Тузовский А. Ф. Обзор подходов семантического поиска //Доклады Томского государственного университета систем управления и радиоэлектроники. – 2010. – №. 2-2 (22).
116. Кузнецов О. П., Суховеров В. С., Шипилина Л. Б. Онтологии в современных информационных системах //Датчики и системы. – 2011. – №. 8. – С. 67-77.
117. Апанович З. В., Винокуров П. С., Кислицина Т. А. Средства визуального анализа информационного наполнения порталов, входящих в облако Linked Open Data //Труды. – 2011. – С. 113-120.
118. Suárez-Figueroa M. C., D'Aquin M., Kronberger G. Combining data mining and ontology engineering to enrich ontologies and linked data. – 2012.
119. Zhao L., Ichise R. Mid-ontology learning from linked data //Joint International Semantic Technology Conference. – Springer, Berlin, Heidelberg, 2011. – С. 112-127.
120. Jain P. et al. Ontology alignment for linked open data //International semantic web conference. – Springer, Berlin, Heidelberg, 2010. – С. 402-417.
121. Оробинская Е. А., Дорошенко А. Ю. Использование онтологий для автоматической обработки текстов на естественном языке. – 2011.
122. Добров Б. В., Лукашевич Н. В. Тезаурус РуТез как ресурс для решения задач информационного поиска //Труды Всероссийской Конференции Знания-Онтологии-Теории (ЗОНТ-09), Новосибирск. – 2009. – Т. 10.
123. Ле Хоай, Тузовский, А.Ф.: Разработка семантических электронных библиотек на основе онтологических моделей. Труды XV Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2013, г. Ярославль, 14 – 17 октября 2013 года, сс. 143- 151 (2013)
124. Ding L. et al. owl: sameAs and Linked Data: An empirical study //Proceedings of the Second Web Science Conference. – 2010.
125. Volz J. et al. Discovering and maintaining links on the web of data //International Semantic Web Conference. – Springer, Berlin, Heidelberg, 2009. – С. 650-665.
126. Ngonga Ngomo A. C. et al. Sorry, i don't speak SPARQL: translating SPARQL queries into natural language //Proceedings of the 22nd international conference on World Wide Web. – ACM, 2013. – С. 977-988.

127. Lihua Z. et al. Ontology Integration for the Linked Open Data. – 2013. // https://ir.soken.ac.jp/?action=repository_action_common_download&item_id=4906&item_no=1&attribute_id=19&file_no=2
128. 2010 Mathematics Subject Classification // https://mathscinet.ams.org/msc/msc2010.html
129. Lange C. et al. Reimplementing the mathematics subject classification (msc) as a linked open dataset //International Conference on Intelligent Computer Mathematics. – Springer, Berlin, Heidelberg, 2012. – С. 458-462.
130. Elizarov A. M. et al. Mathematical knowledge representation: semantic models and formalisms //Lobachevskii Journal of Mathematics. – 2014. – Т. 35. – №. 4. – С. 348-354.
131. Сытник А. А., Вагарина Н. С., Мельникова Н. И. Онтологическое описание мультимедийных ресурсов в контексте технологий семантического веб //Вестник Саратовского государственного технического университета. – 2011. – Т. 4. – №. 2 (60).
132. Сытник А. А., Вагарина Н. С., Мельникова Н. И. Онтологическое описание мультимедийных ресурсов в контексте технологий семантического веб //Вестник Саратовского государственного технического университета. – 2011. – Т. 4. – №. 2 (60).
133. Сотников А. Н. и др. Принципы построения и формирования электронной библиотеки «Научное наследие России» //Программные продукты и системы. – 2012. – №. 4.
134. Roknuzzaman M., Kanai H., Umemoto K. Integration of knowledge management process into digital library system: a theoretical perspective //Library Review. – 2009. – Т. 58. – №. 5. – С. 372-386.
135. Nielsen U., Eriksson P. The integration of digital library services in an academic environment. – 2012.
136. Ломов П. А., Шишаев М. Г., Диковицкий В. В. Упрощенное представление OWL-онтологий для их применения в графических пользовательских интерфейсах //Труды Кольского научного центра РАН. – 2012. – Т. 3. – №. 4.
137. Городецкий В. И., Тушканова О. Н. Онтологии и персонификация профиля пользователя в рекомендующих системах третьего поколения //Онтология проектирования. – 2014. – №. 3 (13).
138. Циркин Б.Г. Использование онтологического подхода к разработке каталога пользовательских предпочтений. — RCDL. - Дубна, Объединенный институт ядерных исследований. - 2014. - С. 145- 148.
139. Елизаров А. М. и др. Онтологии математического знания и рекомендательная система для коллекций физико-математических документов //Докл. РАН. – 2016. – Т. 467. – №. 4. – С. 392-395.
140. Соколов А. В. Что есть информационная потребность? //Труды Санкт-Петербургского государственного института культуры. – 2013. – Т. 197.
141. Шокин Ю. И., Федотов А. М., Барахнин В. Б. Проблемы поиска информации. – 2010.
142. Nnadi N., Bieber M. 1Towards Lightweight Digital Library Integration. – 2004.
143. Васильев А. В. Шаблон проектирования корпоративных Java-приложений, построенных на основе адаптивных моделей данных, обеспечивающий их масштабируемость //Труды Московского физико-технического института. – 2013. – Т. 5. – №. 4 (20).
144. Alani H., Harris S., O'Neill B. Ontology winnowing: A case study on the akt reference ontology //International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06). – IEEE, 2005. – Т. 2. – С. 710-715.
145. DBLP computer science bibliography // https://dblp.uni-trier.de/
146. Autonomous citation indexing and literature browsing using citation context // https://citeseerx.ist.psu.edu
147. Community Research and Development Information Service // https://cordis.europa.eu/en
148. The Engineering and Physical Sciences Research Council // https://epsrc.ukri.org/
149. Association for Computing Machinery // https://www.acm.org/

150. Institute of Electrical and Electronics Engineers // https://www.ieee.org/
151. Lynch C. A. The Z39. 50 information retrieval standard //D-lib Magazine. – 1997. – T. 3. – №. 4.
152. Hammer S., Favaro J. Z39. 50 and the world wide web //D-Lib magazine. – 1996. – №. March.
153. Gonçalves M. A., Fox E. A., Watson L. T. Towards a digital library theory: a formal digital library ontology //International Journal on Digital Libraries. – 2008. – T. 8. – №. 2. – C. 91-114.
154. Latif A. et al. Discovery and Construction of Authors' Profile from Linked Data (A case study for Open Digital Journal) //LDOW. – 2010. – T. 29. – C. 1-4.
155. Latif A., Scherp A., Tochtermann K. LOD for library science: benefits of applying linked open data in the digital library setting //KI-Künstliche Intelligenz. – 2016. – T. 30. – №. 2. – C. 149-157.
156. Latif A. et al. Turning keywords into URIs: simplified user interfaces for exploring linked data //Proceedings of the 2nd international conference on interaction sciences: information technology, culture and human. – ACM, 2009. – C. 76-81.
157. Dou D., Wang H., Liu H. Semantic data mining: A survey of ontology-based approaches //Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015). – IEEE, 2015. – C. 244-251.
158. Binding C., Tudhope D. Improving interoperability using vocabulary linked data //International Journal on Digital Libraries. – 2016. – T. 17. – №. 1. – C. 5-21.
159. Bernard J. et al. Visinfo: a digital library system for time series research data based on exploratory search—a user-centered design approach //International Journal on Digital Libraries. – 2015. – T. 16. – №. 1. – C. 37-59.
160. Börner K. et al. VIVO: A semantic approach to scholarly networking and discovery //Synthesis lectures on the Semantic Web: theory and technology. – 2012. – T. 7. – №. 1. – C. 1-178.

ПРИЛОЖЕНИЕ

Документы, удостоверяющие практическое использование результатов диссертационного исследования

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2019615657

«LibMeta-LOD-ресурс»

Правообладатель: ***Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (RU)***

Авторы: ***Атаева Ольга Муратовна (RU), Серебряков Владимир Алексеевич (RU), Теймуразов Кирилл Борисович (RU)***

Заявка № **2019614064**
Дата поступления **16 апреля 2019 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***06 мая 2019 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2019615409

«LibMeta-микротезаурус»

Правообладатель: ***Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (RU)***

Авторы: ***Атаева Ольга Муратовна (RU), Серебряков Владимир Алексеевич (RU), Тучкова Наталья Павловна (RU)***

Заявка № **2019614050**
Дата поступления **16 апреля 2019 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***26 апреля 2019 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2019615406

«LibMeta-онтология»

Правообладатель: ***Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (RU)***

Авторы: ***Атаева Ольга Муратовна (RU), Серебряков Владимир Алексеевич (RU), Заварза Георгий Николаевич (RU)***

Заявка № **2019614061**
Дата поступления **16 апреля 2019 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***26 апреля 2019 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2019615658

«LibMeta-рекомендательная подсистема»

Правообладатель: ***Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (RU)***

Авторы: ***Атаева Ольга Муратовна (RU), Серебряков Владимир Алексеевич (RU), Меденников Антон Михайлович (RU), Босов Алексей Вячеславович (RU)***

Заявка № **2019614062**
Дата поступления **16 апреля 2019 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***06 мая 2019 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

Г.П. Ивлиев

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2019615408

«LibMeta-тезаурус»

Правообладатель: ***Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (RU)***

Авторы: ***Атаева Ольга Муратовна (RU), Серебряков Владимир Алексеевич (RU), Тучкова Наталья Павловна (RU)***

Заявка № **2019614054**
Дата поступления **16 апреля 2019 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***26 апреля 2019 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2019615407

«Конструктор LibMeta-онтология»

Правообладатель: ***Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (RU)***

Авторы: ***Атаева Ольга Муратовна (RU), Серебряков Владимир Алексеевич (RU), Босов Алексей Вячеславович (RU)***

Заявка № **2019614060**
Дата поступления **16 апреля 2019 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***26 апреля 2019 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев