Организация анализа научных данных на основе семантических технологий

М.Ш. Муртазина[a], Т.В. Авдеенко[a]

[a] *Новосибирский государственный технический университет, 630073, пр. Карла Марскса, 20,Новосибирск, Россия*

Аннотация

В статье исследуется существующая в мире структура знаний в научной сфере, организованная средствами Semantic Web. Дается обзор технологий для описания семантики, средств построения запросов к семантическим данным, проектов публикации связанных открытых данных (Linked Open Data, LOD). На основании анализа сложившейся ситуации в мире установлено, что в России процесс создания структур знаний, обеспечивающих доступность семантической информации в области научных решений и технологий в сравнении с рядом зарубежных стран, находится на начальном этапе. Предлагается подход к организации онтологической базы знаний, опирающейся на базовые принципы LOD, для интеллектуальной поддержки научно-исследовательской деятельности.

Ключевые слова: информатизации научно-исследовательской деятельности; интеллектуальный анализ данных; онтология; открытые связанные данные; открытые научные знания; Semantic Web

1. Введение

По объемам данных World Wide Web и множество других систем передачи данных в сети Интернет являются самым большим из когда-либо существовавших в истории человечества хранилищ данных. Развитие науки, производства и других отраслей человеческой деятельности в последние десятилетия сопровождалось накоплением больших объемов данных в сети Интернет. Этот процесс накопления включал в себя создание банков данных, систематизацию информации по определенному набору атрибутов. С начала 2000-х годов появилась новая задача – поиска данных в пределах имеющихся баз данных с учетом семантики. Данная задача может быть эффективна решена по средствам построения и применения онтологий. В настоящее время существует ряд крупномасштабных проектов по созданию онтологических баз знаний, среди которых наиболее известны DBpedia [1], Wikidata [2], FOAF [3], GeoNames [4],YAGO2 [5] и Bio2RDF [6].

Буквально в последние несколько лет рядом крупных поисковых систем внедряется технология семантического поиска. Так, в 2012 году в поисковую систему Google [7] был добавлен инструмент семантического поиска Knowledge Graph, который не просто идентифицирует ключевые слова, но и сопоставляет с базой фактов об искомом объекте. Данный инструмент показывает отличные результаты в области поиска мест, людей и предметов, но при поиске информации по научно-технологическим решениям, технологиям и производствам в большинстве случаев по прежнему возвращает результаты соответствующие исключительно вхождению ключевых слов.

Задача поиска и оценки полезности научных публикаций посредствам семантики, извлечения знаний из найденных данных является актуальной в мировой науке. Ярким примером тому является запуск в 2015 году проекта Semantic Scholar [8], который представляет собой систему поиска по научным публикациям с элементами искусственного интеллекта. Алгоритмы сортировки результатов поиска данной системы учитывают научную значимость работ, которая определяется по их цитируемости, но не путем простого подсчета количества цитат, а путем определения контекста, в котором встречается упоминание работы (обзор литературы, описание методологии научного исследования и т.д.), также учитываются используемые речевые обороты при отсылке на публикацию. На текущий момент указанная система выполняет анализ по англоязычным публикациям.

В последние пять лет потребность в создании методов и моделей поддержки принятия решений на основе анализа научных данных в России выражается в грантах и проектах различных фондов. Таким образом, задача создания модели информационного пространства в области научных данных, которая позволит эффективно использовать данные, собранные с открытых русскоязычных онлайн источников, представляется актуальной. Решение этой задачи требует создания онтологической базы знаний, содержащей информацию из множества источников в области научных и технологических решений, которая может применяться в процессе принятия решений.

В настоящей работе предлагается подход к организации онтологической базы знаний, опирающейся на базовые принципы LOD, для интеллектуальной поддержки научно-исследовательской деятельности. Статья организована следующим образом. В разделе 1 обосновывается актуальность темы исследования. В разделе 2 анализируется понятийный аппарат. В разделе 3 дается обзор семантических технологий и проектов публикации связанных открытых данных. В разделе 4 дается оценка текущего состояния решения задача поиска по русскоязычным научным публикация. В разделе 5 представлен обзор существующих онтологий для описания научных публикаций и процесса научно-издательской деятельности. В разделе 6 предлагается подход к организации онтологической базы знаний в области научной деятельности. В разделе 7 делаются выводы о перспективах применения семантических технологий при информатизации научно-исследовательской деятельности.

2. Данные для научных исследований и научные данные

Термин «данные» тесно связан и с такими общеизвестными терминами, как «информация» и «знания». В 1989 году известный американский учёный Рассел Акофф в работе «От данных к мудрости» [9] определил иерархию типов содержания человеческого разума: данные – информация – знание – понимание – мудрость. По Р.Акофу, данные – это «символы, представляющие свойства объектов, событий и их окружения». Данные – это результаты наблюдений. Информация выводится из данных, и состоит из описаний, ответов на вопросы: кто, что, где, когда, сколько. Знание позволяет преобразовать информацию в инструкции, обнаруживая связи между компонентами информации. Понимание – осознание закономерностей в знаниях, мудрость – «способность оценивать любой выбор по степени прогресса в продвижении к этому метаидеалу».

В рамках задачи информатизации научно-исследовательской деятельности целесообразно разграничить такие понятия, как «данные для научных исследований» и «научные данные».

Данные для научных исследований – это исходные для исследования сведения, выраженные как в числовой, так и любой другой форме.

Научные данные – это данные, представляющие собой научные результаты исследований. Под понятием «научным результат» («научно-технический результат») в соответствии со статьей 2 Федерального закона от 23 августа 1996 г. N 127-ФЗ «О науке и государственной научно-технической политике» понимается «продукт научной и (или) научно-технической деятельности, содержащий новые знания или решения и зафиксированный на любом информационном носителе» [10].

Совокупность научных данных в определенной предметной области образует научные знания. Научные данные по своей природе уже являются знаниями, так как – это результат обработки исходных данных о сущностях предметной области, их свойствах и отношениях между ними. Иными словами научные знания – это систематизированные обобщенные знания.

Выделяется два базовых источника знаний – специалисты предметной области и лингвистический корпус текстов, под которым понимается коллекция текстов по заданной тематике[11]. Источниками научных знаний соответственно являются ученые и написанные ими научные труды, большая часть которых доступна в электронной форме.

На современном этапе развития данных стало настолько много (даже появилось устойчивое выражение «большие данные»), что данные сами по себе перестали представлять большую ценность. То, что действительно ценно – это знания (появился даже термин «большие знания»), которые можно извлечь из данных. В этой связи постоянно возрастающий объем научных публикаций, размещаемых на информационных ресурсах сети Интернет, требует организации поиска информации, релевантной запросу пользователя. Это может быть достигнуто благодаря семантическим технологиям.

3. Семантические технологии

В 1989 году Тим Бернерс-Ли предложил концепцию Всемирной паутины (World Wide Web, WWW), в 1998 – концепцию Семантической паутины (Semantic Web), а в 2007 году – дальнейшую концепцию развития Гигантский глобальный граф (Giant Global Graph).

Семантические технологии – это информационные технологии управления знаниями, которые основываются на моделях представления данных в семантической форме. Цель семантических моделей представления данных заключается в обеспечении возможности записи информации в форме, которая может быть обработана компьютером с учетом смыслового значения информации. Семантические технологии были названы аналитиками Gartner одним из наиболее многообещающих ИТ-трендов 2013 года [12].

К основным семантическим технологиям относятся:

- RDF (Resource Description Framework) – язык записи триплетов, основанных на модели «субъект-предикат-объект»;
- RDFS (Resource Description Framework Schema)– язык описания схем RDF;
- OWL (Web Ontology Language) – язык описания онтологий;
- SPARQL (Protocol and RDF Query Language) – технология создания хранилищ RDF-данных, а также язык запросов для извлечения данных из RDF.

В настоящее время сематические модели и технологии активно используются в самых разнообразных системах управления данными для обеспечения извлечения знаний из связанных данных (Linked Data, LD). Связанными данными называют взаимосвязанные наборы данных в World Wide Web, а также метод публикации структурированных данных, обеспечивающий доступ к семантическому описанию, и, соответственно, возможность построения семантических запросов к публикуемым данным. Объекты связанных данных представляют собой распределенные объекты данных, имеющие стилистически единообразные идентификаторы, т.е. данные должны быть представлены в формате пригодном для разыменования URI [13]. Чаще всего для этого применяется стандартизированная модель описания объектов данных для Semantic Web – RDF.

Термин связанные данные (Linked Data) был введен в научный оборот в 2006 году, а чуть позже в начале 2007 года появился и термин открытые связные данные (Linked Open Data, LOD). С этого времени активно развивается разработка интеллектуальных информационных систем, извлекающих знания из LOD-облаков и данных, полученных с помощью комбинации статистических и лингвистических методов.

LOD-облако – это модель онлайн-хранилища, в котором содержатся наборы связанных данных, доступных пользователям сети Интернет.

В январе 2007 стартовал проект «Linking Open Data Project», который представляет из себя LOD-облако из LOD-облаков. В рамках проекта каждое облако рассматривалось как набор данных. В мае 2007 года LOD-облако проекта включало в себя 12 наборов данных, к ноябрю 2007 количество наборов данных увеличилось до 25. Только за первые пять лет существования проекта LOD-облако охватывало более 50 миллиардов фактов из самых разных областей, таких как география, средства массовой информации, биология, химия, экономика, энергетика и т.д. [14]. Эти данные были различного качества и большинство из них могло быть повторно использовано в коммерческих целях, в том числе и при разработке систем поддержки принятия решений. В настоящее время LOD-облако включает 570 наборов данных [15]. Ядром LOD-облака является проект DBpedia, который в свою очередь тоже является LOD-облаком.

DBpedia — это краудсорсинговый проект, направленный на извлечение структурированной информации, собранной в рамках проекта свободной энциклопедии «Википедия». В Википедии для структурированной различных типов информации применяется вики-разметка, которая включает в себя шаблоны-инфобоксы, изображения, гео-координаты, ссылки на внешние веб-страницы, ссылки на различные языковых версий страницы. Фреймворк DBpedia извлекает эту структурированную информацию из Википедии и представляет пользователям в виде базы знаний. DBpedia позволяет задавать сложные запросы к Википедии, а также подсоединять различные наборы данных в Интернете к данным Википедии. На текущий момент английская версия базы знаний DBpedia описывает свыше 4,58 миллион фактов, из которых 4,22 миллионов организованы в согласованные онтологии предметных областей. Кроме того DBpedia предлагает локализованные версии на 125 языках. Наборы данных DBpedia связаны с другими LOD-наборами данных примерно 50 миллионами RDF-ссылок [1].

Основными преимуществами открытых связных данных являются их доступность и машиночитаемость. В рамках реализации идеи об открытых данных по всему миру реализуются правительственные инициативы «Открытые данные государств» [16]. 26 марта 2014 года был запущен федеральный правительственный Портал открытых данных в России [17]. Проекты открытых данных реализуются и на региональном уровне. В частности, в России 29 января 2013 года был запущен проект «Портал открытых данных Правительства Москвы» [18]. В последние годы стала актуальна задача анализа LOD-наборов данных для журналистики данных, LOD стали широко применяться в рекомендательных системах с целью повышения эффективности их работы. За последние несколько лет публикация LOD-наборов данных для науки стала перспективным направлением информатизации научно-исследовательской деятельности в мире. И здесь следует отметить, что хотя концепции открытых данных для науки (Open Data In Science) появилась еще в 1950-х годах [19], ее реализация действительно стала возможной именно с появлением семантических технологий представления и поиска данных.

4. Характеристика информационных ресурсов, предоставляющие услуги поиска по русскоязычным научным публикациям

Все информационные ресурсы, предоставляющие услуги поиска по научным публикациям, могут быть условно разделены на несколько групп:

1) научные поисковые каталоги, содержащие данные о научных публикациях, размещенных в сети Интернет;

2) информационные системы с веб-интерфейсом, предоставляющие пользователям возможность поиска по базам данных этих информационных систем по определенному набору параметров;

3) социальные сети ученых;

4) сайты, содержащие научные публикации по результатам исследований, но не содержащие встроенные средства поиска.

К представителям первой группы можно отнести ряд международных и зарубежных проектов, которые поддерживают индексирование русскоязычных источников, таких как GoogleScholar [20] и ScienceDirect [21]. Среди отечественных разработок – это поисковая система научных публикаций Scholar.ru [22]. Проекта Scholar.ru представляет собой каталог ссылок на научные работы, размещенные на различных ресурсах сети Интернет. Основной цель проекта Scholar.ru является сбор информации о свободно скачиваемых научных публикациях. Поисковый механизм системы позволяет искать публикации по разделу (по областям наук), автору, учреждению, названию журнал, году публикации (задается период для поиска), URL сайта, на котором размещена публикация.

Ко второй группе относятся такие ресурсы, как информационная система Российского фонда фундаментальных исследований [23], База данных технологий научно-технического форума [24], Интернет-портал RSCI.RU [25], сайт Российского гуманитарного научного фонд [26], сайт Российского научного фонд [27] и т.д. Проведенный анализ поисковых форм на ресурсах второй группы показал, что механизмы поиска значительно ограничены: не учитывается морфология слов, отсутствует возможность поиска информации в контексте и не учитывается семантика слов. В большинстве случаев в поиске задействованы следующие атрибуты:

1) название и/или автор(-ы) работы;

2) область знаний (по областям наук);

3) год издания работы (или заявки на проведение научных исследований).

Среди ресурсов данной группы следует выделить научные электронные библиотеки (например, eLIBRARY.RU [28]) и системы наукометрических данных, такие как интеллектуальная информационная системам «Истина» [29]. В

системе «Истина», разработанной в МГУ, используются онтология, методы лингвистического и статистического анализа текстов, контекстный анализ результатов поиска в Интернет [30, с.137].

Примером ресурсов третьей группы являются Scipeople [31], Социальная сеть «Учёные России» [32] и Социальная научная сеть Scientific Social Community [33].

К четвертой группе относятся ресурсы, на которых размещаются архивы номеров журналов, материалы конференций, объявления о конкурсах научных проектов, отчеты по результатам научным исследований, но отсутствуют формы поиска по данным материалам или есть только поиск по страницам сайта по введенному фрагменту текста. Например, сайт института экономики Российской академии наук [34] и сайт фонда перспективных исследований [35]. Поиск информации (с точки зрения пользователя ресурса) по четвертой группе ресурсов представляет очень трудоёмкую задачу, поскольку данные не структурированы. Тексты научных публикаций и отчетов по результатам научных исследований могут быть просто прикреплены на веб-странице в форме текстовых (pdf, doc(x)), а порой и графических файлах.

Анализ информационных ресурсов сети Интернет, содержащих данные по научным и технологическим решениям, показал, что поиск информации на них организован по принципу нахождения совпадений со значениями полей баз данных этих систем. Кроме того на отдельных сайтах для поиска информации просто подключаются сервисы известных поисковых систем. При описании ресурса модель представления данных RDF зачастую не применяется, и, соответственно, практически нет SPARQL-точек доступа к русскоязычным данным. Доступность семантической русскоязычной информации в области научных решений и технологий в сравнении с рядом зарубежных стран, находится на начальном этапе. Таким образом, задача создание модели онтологической базы знаний в области научной информации, связной с открытыми источниками научной информации, включая LOD-облака, и инструментария для извлечения знаний из внешних источников является актуальной задачей.

5. Обзор существующих онтологий для описания научных публикаций

На сегодняшний день разработано несколько онтологий для описания научных публикаций и процесса научно-издательской деятельности: BIBO[36], комплекс онтологий SPAR [37], CERIF [38], SWRC[39], EXPO [40], FRBR [41], SKOS [42], Dublin Core [43], ЕНИП [44] и др. Далее рассмотрим области применения перечисленных онтологий.

Онтология BIBO включает в себя основные понятия и свойства для описания библиографических ссылок на Semantic Web в RDF (т.е. цитаты, книги, статьи и т.д.).

Комплекс онтологий SPAR позволяет описать процесс публикации с помощью RDF. Базовые онтологии набора:

– FaBiO (FRBR-aligned Bibliographic Ontology) – онтология, позволяющая описывать библиографические объекты (журнальные статьи, материалы конференций, книги и т.д.), которые содержат библиографические ссылки;

– CiTO (Citation Typing Ontology) – онтология, предназначенная для описания природы цитат в научных публикациях (факт или утверждение);

– BiRO (Bibliographic Reference Ontology) – онтология, предназначенная для описания библиографических записей и ссылок, и их компиляцию в библиографические сборники и библиографические списки;

– C4O (Citation Counting and Context Characterisation Ontology) - онтология, которая позволяет оценивать цитаты из цитируемых источников по их числу и расположению в контакте;

– DoCO (Document Components Ontology) – онтология, которая содержит структурированный словарь компонентов документа, включает структурные блоки (например, параграф, раздел, глава) и функциональные блоки (например, введение, обсуждение, благодарность, список литературы, рисунок, приложение);

– PSO (Publishing Status Ontology) – онтология, которая предназначена для описания состояния публикации на каждом этапе издательского процесса;

– PRO (Publishing Roles Ontology) – онтология, характеризующая роли агентов - людей, юридических лиц и вычислительных средств в процессе публикации. Агентами могут быть автор, редактор, рецензент, издатель или библиотекарь;

– PWO (Publishing Workflow Ontology) – онтология для описания шагов в рабочих процессах, связанных с публикацией документа.

CERIF (Common European Research Information Format) – онтология для описания процесса научно-исследовательской деятельности. На верхнем уровне расположены сущности «Персона», «Проект», «Организационный блок», которые связаны с сущностями других уровней, например с сущностями «Публикация», «Продукт», «Патент».

Semantic Web for Research Communities (SWRC) - онтология для моделирования объектов исследовательских сообществ, таких как персоны, организации, публикации (библиографические метаданные) и их отношений.

EXPO (EXPeriment) - онтология для описания научных экспериментов, включающая около 200 концептов.

FRBR (Functional Requirements for Bibliographic Records) - онтология, позволяющая описывать библиографические записи. Разделена на три группы: первая группа позволяет описать результаты интеллектуального труда (работа, выражение, манифестация, экземпляр), вторая – лицо, ответственное за результат интеллектуального труда (персона, группа лиц, юридическое лицо), третья – включает сущности, связанные с первой и второй группами (понятие, объект, событие, место).

SKOS (Simple Knowledge Organization System) - онтология для описания тезаурусов по модели RDF.

Dublin Core - онтология, включающая два набора метаданных: простой и расширенный. Первый состоит из 15 элементов, второй - из 22 классов и 55 элементов. Спецификация Dublin Core имеет статус официального международного стандарта ISO 15836:2009 [45].

ЕНИП (Единое Научное Информационное Пространство) - проект, нацеленный на интеграцию научных данных учреждений РАН. Онтология, используемая в проекте, основана на онтологии Dublin Core. В онтологии ЕНИП выделяется четыре основные группы сущностей (участники научной деятельности, научная деятельность, результаты научной деятельности, документы и публикации).

Проведенный анализ существующих онтологий показывает, что достаточно хорошо проработаны онтологии для описания домена «библиографические сущности», однако таких онтологий в ряде случае не достаточно для анализа научных данных. Например, если необходимо проанализировать востребованность решения некоторой научной проблемы необходимо не только анализировать уже имеющие публикации с описание научных результатов, но и получать срез знаний о текущих потребностях. Указанные знания могут быть извлечены из данных об подаваемых грантах, а также проектах различных фондов. В следующей части работы представляется модель, учитывающая научные события.

6. Модель онтологической базы знаний в области научно-исследовательской деятельности

Модель онтологической системы в общем виде может быть представлена следующим образом:

$$Z = <O_m, O_p, M> \quad (1)$$

где O_m – онтология верхнего уровня (метаонтология), которая включает в себя наиболее абстрактные термины и отношения между этими терминами;

O_p – множество предметных онтологий;

M –механизм вывода знаний.

Представляется, что модель онтологической базы знаний в области научно-исследовательской деятельности, может быть основана на мета-онтологии, приведенной на рис. 1.

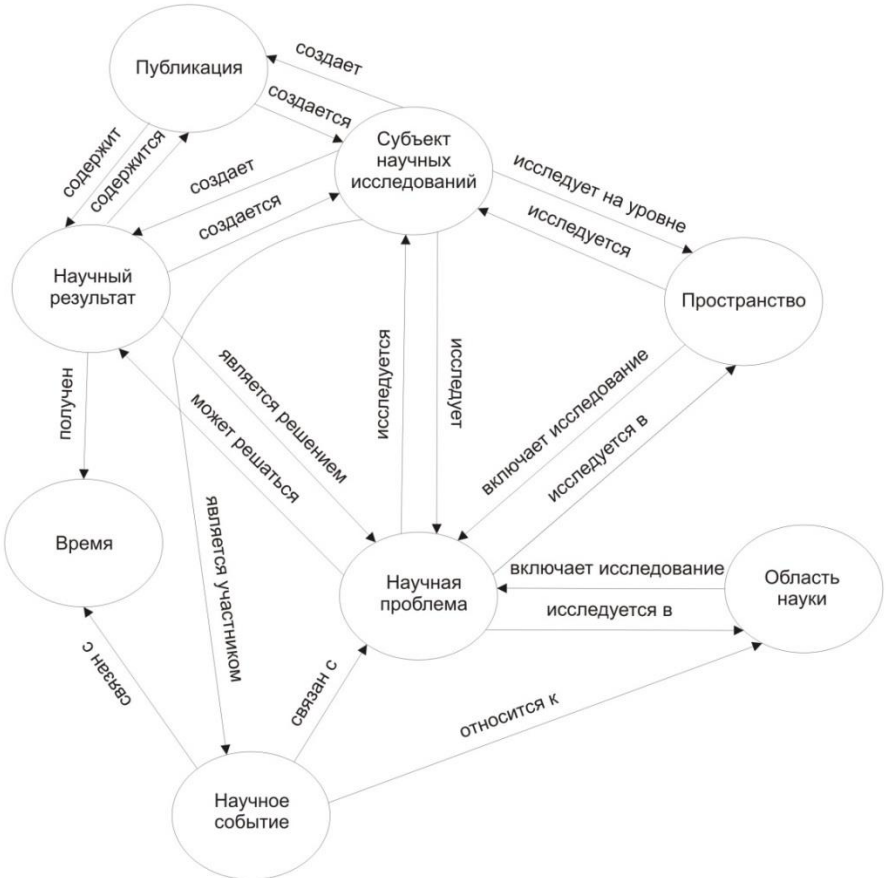

Рис. 1. Онтология верхнего уровня.

Каждый узел представленной сети является вершиной классовой иерархии. Примеры подклассов для классов онтологии верхнего уровня приведены в таблице 1.

Таблица 1.Примеры подклассов для классов онтологии верхнего уровня

Класс	Подкласс
субъект научных исследований	персона, научный коллектив, организация и т.п.
научный результат	исходная попытка, оценка состояния, теория, парадигма, инструмент, модель и т.д.
пространство	географическая локация, политическое пространство, культурное пространство, природное пространство и т.д.
время	год, квартал, месяц, период и т.д.
область науки	классификаторы УДК, ГРНТИ, РФФИ и т.д.
научная проблема	субстратная проблема, структурная проблема и т.д.
научное событие	объявление о конкурсе, конференция, форум и т.д.

Онтологии нижележащих уровней могут быть заданы как онтологии вида [46]:

$$O_F =< C, R, S, G, T, D_S, D_G, E > \quad (2)$$

где $C = \{c_i \mid i = 1,...n\}$ – конечное непустое множество классов, описывающих понятия предметной области;

$R = \{r_i \mid i = 1,...m\}$ – конечное множество бинарных отношений, заданных на классах, $R \subseteq C \times C$, $R = \{R_{ISA}\} \cup R_{ASS}$, где R_{ISA} – антисимметричное, транзитивное, нерефлексивное отношение иерархии «класс-подкласс», задающее частичный порядок на множестве классов; R_{ASS} – конечное множество ассоциативных отношений;

$S = \{s_i \mid i = 1,...k\}$ – конечное множество слотов (атрибутов класса);

$G = \{gs_i \mid i = 1,...l\}$ – конечное множество фасетов (атрибутов слота);

T – конечное непустое множество, определяющее контролируемый словарь терминов предметной области, построенное на множестве базовых терминов $B = \{b_i \mid i = 1,...n\}$, составляющих множество имен классов онтологии:

$$T = \bigcup_{i=1}^{n} T_i, T_i = \{b_i\} \cup Eq(b_i), \quad \bigcap_{i=1}^{n} T_i = \varnothing;$$

$Eq(b_i)$ – множество синонимичных терминов, каждый из которых связан с базовым термином $b_i \in B$;

D_S – конечное множество типов слотов;

D_G – конечное множество типов фасетов;

$E = \{e_i \mid i = 1,...u\}$ – конечное множество экземпляров классов.

Структура класса определяется следующим образом:

$$c =< Name_C, (isa\ c_{parent}), (s_1, s_2, ...s_{n(c)}) > \quad (3)$$

где c, $c_{parent} \in C$ – классы онтологии, связанные отношением иерархии R_{ISA};

$s_i \in S$ – слоты,

$Name_c \in B$ – имя класса, являющееся базовым термином контролируемого словаря T.

Иерархии классов образуются посредством указания в подчиненном классе связи «is-a» и имени класса-родителя c_{parent}.

Все множество классов C разбивается на два непересекающихся подмножества $C = C_{abstract} \cup C_{concrete}$. Для классов подмножества $C_{concrete}$ возможно определять экземпляры класса (конкретные объекты) $e \in E$. Структура класса-экземпляра аналогична структуре класса c, для которого построен экземпляр:

$$e(c) =< Name_e, (s_1^e, s_2^e, ..., s_{n(c)}^e) > \quad (4)$$

где $s_1^e, s_2^e, ..., s_{n(c)}^e$ – экземпляры слотов класса c, заполненные конкретными значениями атрибутов.

Определение ассоциативных отношений, составляющих множество R_{ASS}, осуществляется путем явного указания в качестве значения слота имени связанного с ним класса, а также типа связи, существующей между этими классами. Для

реализации ассоциативных связей среди элементов множества типов слотов D_S кроме выделения подмножества стандартных типов (symbol, string, float,…) D_{SS}, используются также типы D_{class} (тип «Класс») и $D_{instance}$ (тип «Экземпляр»):

$$D_S = D_{SS} \cup \{D_{class}\} \cup \{D_{instance}\}.$$

Задание типов D_{class} и $D_{instance}$ предполагает указание дополнительного аргумента – ассоциированного класса. Если один из слотов класса c_1 имеет тип $D_{instance}$ с ассоциированным классом c_2, то в качестве значений слота при создании экземпляров класса c_1 могут быть использованы экземпляры классов множества $Tr(c_2)$ – транзитивного замыкания c_2 по отношению R_{ISA}, включающего класс c_2 и все его подклассы ниже по иерархии:

$$Tr(c_2) = \{c_2\} \cup \{c_i \in C \mid \exists R_{ISA}(c_i, c_2)\}.$$

В этом случае классы c_1 и c_2 связаны ассоциативным отношением, т.е. $\exists R_{ASS}(c_1, c_2)$. Если один из слотов класса c_1 имеет тип D_{class} с ассоциированным классом c_2, то в качестве значений слота при создании экземпляров класса c_1 могут быть использованы классы множества $Tr(c_2)$. В этом случае классы c_1 и c_2 также связаны ассоциативным отношением, т.е. $\exists R_{ASS}(c_1, c_2)$. Таким образом, значением слота может становиться не только экземпляр ассоциированного класса, но и базовый термин, что может использоваться для описания сложных объектов предметной области терминами контролируемого словаря.

Структура слота определяется следующим образом:

$$s_C = < Name_{S,C}, (gs_1, gs_2, \ldots gs_{k(S,C)}) > \quad (5)$$

где $s_c \in S$ – слот класса c;

$gs_i \in G$ – фасеты слота;

$Name_{s,c}$ – имя слота.

Семантические связи между отдельными областями могут быть описаны по стандартизированной модели описания объектов данных для Semantic Web – RDF, что позволяет соединять имеющие в машиночитаемом формате данные. Для реализации модели использован редактор Protégé. Данный редактор позволяет подключать к проекту внешние OWL-онтологии, что обеспечивает возможность использования уже существующих онтологий в качестве составных частей разрабатываемой онтологии.

На рисунке 2 представлен фрагмент онтологии с такими тестовыми экземплярами классов, как автор, статья и рубрика.

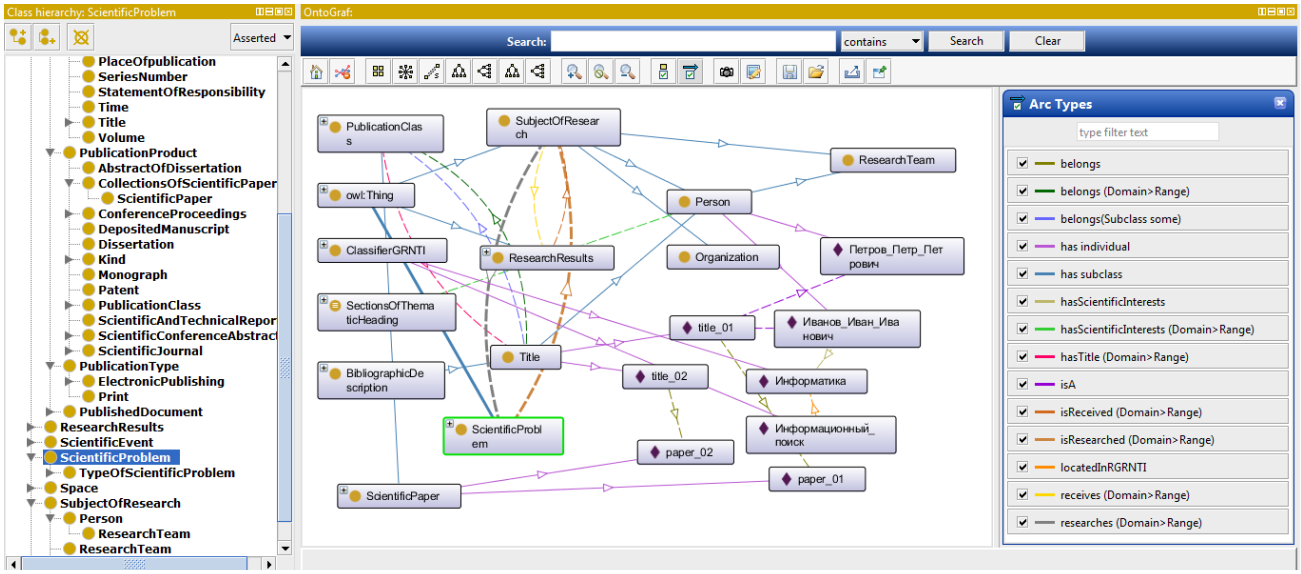

Рис. 2. Фрагмент онтологии.

Одним из ключевых понятий в онтологии является понятие «Публикация», под которым подразумеваются научные публикаций нескольких типов. Разработка первой версии иерархии классов данного понятия «научное издание» произведена на основании ГОСТ 7.60-2003 и ГОСТ Р 7.83-2013. Согласно ГОСТ 7.60-2003, издание – «это документ, предназначенный для распространения содержащейся в нем информации, прошедший редакционно-издательскую обработку, самостоятельно оформленный, имеющий выходные сведения». По признаку физической характеристики издания подразделяются на печатные и электронные. Электронные издания в свою очередь подразделяются на самостоятельные электронные издания, деривативные электронные издания, электронные копии изданий. Для отображения данного деления созданы подклассы “Print” и “ElectronicPublishing”.

Издания могут быть подразделены на опубликованные и неопубликованные. К первой группе относятся такие издания, как монография, сборник научных трудов, материалы конференции, съезда, симпозиума, автореферат диссертации, препринт, авторское свидетельство или патент. Ко второй группе - научно-технические отчеты,

диссертации, депонированные рукописи. Для отображения данного деления созданы классы “PublishedDocument” и “UnPublishedDocument”.

Библиографическое описание любой публикации может включать в себя: заголовок, основное заглавие, сведения, относящиеся к заглавию, сведения об ответственности, сведения об издании, место издания, издательство, дату издания, объем, основное заглавие серии, номер выпуска серии. Класс “BibliographicDescription” включает подклассы, описывающие структуру библиографического описания публикации.

Обычно научные публикации сопровождаются метаданными, включающими заголовок, ключевые слова и аннотацию. Для решения задачи организации семантического поиска по определенной проблеме необходимо анализировать содержание этих метаданных. Что может быть достигнуто путем подключения онтологий понятий предметных областей и использования логических условий предметизации в виде продукционных правил.

7. Заключение

В работе показано, что анализ научных данных, представленных на естественном языке, является крайне актуальной задачей в последние десять лет по всему миру. В целях решения данной задачи уже разработано множество онтологий, которые позволяют описывать научные публикации в машиночитаемом формате. Стоит отметить, что одной из ключевых проблем при совмещении онтологий является установка соответствия между классами.

Онтологический подход к описанию научных данных является основой для обеспечения возможности анализа больших объемов научных данных с открытых онлайн источников и извлечения из них научных знаний. Даже элементарное эмпирическое сравнение результатов работы поисковых систем, внедряющих идею «граф знаний» и «гигантский глобальный граф», с результатами работы классических поисковых систем, позволяет говорить об неоспоримых преимуществах первых.

В статье рассмотрены возможности семантических технологий для описания данных и исследована степень применения данных технологий к организации информационных ресурсов, предоставляющих услуги поиска по русскоязычным научным публикациям. Проведенный анализ позволяет говорить о необходимости разработки новых и совершенствовании существующих моделей анализа русскоязычных научных данных.

На основании изучения ряда подходов [29,30] к проектированию онтологий предложена модель онтологической базы знаний в области научной деятельности. Реализована первая ее версия в редакторе Protégé. В дальнейшем планируется развить представленную в статье модель и реализовать прототип веб-приложения, обеспечивающего мониторинг научных данных из открытых источников, на основании разрабатываемой модели.

Благодарности

Работа поддержана грантом Министерства образования и науки РФ в рамках проектной части Госзадания, проект № 2.2327.2017/ПЧ «Интеграция моделей представления знаний на основе интеллектуального анализа больших данных для поддержки принятия решений в области программной инженерии».

Литература

[1] DBpedia [Electronic resource]. — Access mode: http://wiki.dbpedia.org/ (25.01.2017).
[2] Wikidata [Electronic resource]. — Access mode: https://www.wikidata.org/wiki/Wikidata:Main_Page (25.01.2017).
[3] FOAF Vocabulary Specification 0.99 [Electronic resource]. — Access mode: http://xmlns.com/foaf/spec/ (25.01.2017).
[4] GeoNames [Electronic resource]. — Access mode: http://www.geonames.org (25.01.2017).
[5] YAGO [Electronic resource]. — Access mode: http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/ (25.01.2017).
[6] Bio2RDF [Electronic resource]. — Access mode: http://bio2rdf.org/ (25.01.2017).
[7] Google: Knowledge Graph [Electronic resource]. — Access mode: http://searchengineland.com/library/google/google-knowledge-graph (25.01.2017).
[8] SemanticScholar [Electronic resource]. — Access mode: https://www.semanticscholar.org/ (25.01.2017).
[9] Акофф, Р. От данных к мудрости [Электронный ресурс]. — Режим доступа: http://www.nsu.ru/xmlui/bitstream/handle/nsu/9064/01.pdf (25.01.2017).
[10] Федеральный закон от 23 августа 1996 г. N 127-ФЗ «О науке и государственной научно-технической политике» (в ред. 23.05.2016 N 149-ФЗ) [Электронный ресурс]. — Режим доступа: http://www.consultant.ru/document/cons_doc_LAW_11507/ (25.01.2017).
[11] Палагин, А.В. Онтологические методы и средства обработки предметных знаний: монография / А.В. Палагин, С.Л. Крывый, Н.Г. Петренко. – Луганск: изд-во ВНУ им. В. Даля, 2012. – 324 с.
[12] Горшков, С. Введение в онтологическое моделирование: учебное пособие [Электронный ресурс] / С. Горшков; ООО «ТриниДата». Екатеринбург, 2016. – 165 с. — Режим доступа: http://trinidata.ru/files/SemanticIntro.pdf (25.01.2017).
[13] Льюис, Д. Обновление концепций RDF и некоторые онтологии [Электронный ресурс] / Д. Льюис. – Режим доступа: http://www.ibm.com/developerworks/ru/library/x-rdfconcepts/ (25.01.2017).
[14] Bauer, F. Linked Open Data: The Essentials. A Quick Start Guide for Decision Makers [Electronic resource] / F. Bauer, M. Kaltenböck. –Austria, 2012. – 60 p. — Access mode: https://www.semantic-web.at/LOD-TheEssentials.pdf (25.01.2017).
[15] The Linking Open Data cloud diagram [Electronic resource]. — Access mode: http://lod-cloud.net/ (25.01.2017).
[16] DATA.GOV [Electronic resource]. — Access mode: https://www.data.gov/ (25.01.2017).
[17] Портал открытых данных Российской Федерации [Электронный ресурс]. — Режим доступа: http://data.gov.ru/ (25.01.2017).
[18] Портал открытых данных Правительства Москвы [Электронный ресурс]. — Режим доступа: http://data.mos.ru/ (25.01.2017).
[19] Mathae, K. B. The case for international sharing of scientific data:a focus on developing countries / K. B. Mathae, P. F. Uhlir. - Washington: The National Academies Press, 2012. - 164 p.
[20] GoogleScholar [Электронный ресурс]. — Режим доступа: http://scholar.google.ru/ (25.01.2017).

[21] ScienceDirect [Electronic resource]. — Access mode: http://www.sciencedirect.com (25.01.2017).
[22] Scholar.ru [Электронный ресурс]. — Режим доступа: http://www.scholar.ru/ (25.01.2017).
[23] Российский фонд фундаментальных исследований [Электронный ресурс]. — Режим доступа: http://www.rfbr.ru/rffi/ru (25.01.2017).
[24] База данных технологий научно-технического форума [Электронный ресурс]. — Режим доступа: http://www.sciteclibrary.ru/rus/catalog/tecs/ (25.01.2017).
[25] RSCI.RU [Электронный ресурс]. — Режим доступа: http://www.rsci.ru/ (25.01.2017).
[26] Российский гуманитарный научный фонд [Электронный ресурс]. — Режим доступа: http://www.rfh.ru/ (25.01.2017).
[27] Российский научный фонд [Электронный ресурс]. — Режим доступа: http://www.rscf.ru/ (25.01.2017).
[28] eLIBRARY.RU [Электронный ресурс]. — Режим доступа: http://elibrary.ru/ (25.01.2017).
[29] Интеллектуальная Система Тематического Исследования НАукометрических данных «ИСТИНА» [Электронный ресурс]. — Режим доступа: https://istina.msu.ru/ (25.01.2017).
[30] Интеллектуальная система тематического исследования научно-технической информации (ИСТИНА) / С.А. Афонин и др. Под ред. академика В.А. Садовничего. – М.: Издательство Московского университета, 2014. – 262 с.
[31] Scipeople [Электронный ресурс]. — Режим доступа: http://scipeople.ru/ (25.01.2017).
[32] Социальная сеть «Учёные России» [Электронный ресурс]. — Режим доступа: http:// www.russian-scientists.ru (25.01.2017).
[33] Социальная научная сеть Scientific Social Community [Электронный ресурс]. — Режим доступа: https://www.science-community.org/ru (25.01.2017).
[34] Институт экономики Российской академии наук [Электронный ресурс]. — Режим доступа: http://inecon.org/publikaczii/katalog-izdanij-ie-ran.html (25.01.2017).
[35] Фонда перспективных исследований [Электронный ресурс]. — Режим доступа: http://fpi.gov.ru/ (25.01.2017).
[36] Bibliographic Ontology Specification [Electronic resource]. — Access mode: http://bibliontology.com/ (25.01.2017).
[37] Semantic Publishing and Referencing Ontologies [Electronic resource]. — Access mode: http://www.sparontologies.net/ontologies (25.01.2017).
[38] euroCRIS [Electronic resource]. — Access mode: http://www.eurocris.org/ (25.01.2017).
[39] SWRC Ontology [Electronic resource]. — Access mode: http://ontoware.org/swrc (25.01.2017).
[40] EXPO [Electronic resource]. — Access mode: http://expo.sourceforge.net/ (25.01.2017).
[41] FRBR [Electronic resource]. — Access mode: http://www.frbr.org/ (25.01.2017).
[42] SKOS Simple Knowledge Organization System RDF Schema [Electronic resource]. — Access mode: https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html (25.01.2017).
[43] Good Ontologies [Electronic resource]. — Access mode: https://www.w3.org/wiki/Good_Ontologies (25.01.2017).
[44] Захаров, А.А. Логическая модель цифровых библиотек в онтологии ЕНИП / А.А. Захаров, В.И. Филиппов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XI Всероссийской научной конференции RCDL'2009. Петрозаводск: КарНЦ РАН, 2009. С. 32-38.
[45] ISO 15836:2009 Информация и документация. Набор элементов метаданных Dublin Core [Электронный ресурс]. — Режим доступа: http://www.iso.org/iso/ru/home/store/catalogue_tc/catalogue_detail.htm?csnumber=52142 (25.01.2017).
[46] Авдеенко, Т.В. Гибридная модель представления знаний для реализации вывода во фреймовой онтологии / Т. В. Авдеенко, М. А. Бакаев // Научный вестник Новосибирского государственного технического университета. - 2013. - № 3. - С. 84-90.