Федеральное агентство по образованию

Московский Физико-Технический Институт
(Государственный Университет)

Факультет Общей и Прикладной Физики

Специальность «Физик??????????»

Кафедра Системная интеграция и менеджмент

Дипломная работа

Исследование и разработка онтологии геополитической системы Российской Федерации

«К защите допущен»:

Зав. кафедрой профессор, д.ф.-м.н.	_____	Клименко С.В.
Научный руководитель, д.ф.-т.н.	_____	Хорошевский В.Ф.
Рецензент, зав. лаб. ЖЗ ИКЛ, д.ф.-м.н.	_____	Сидоров С.С.
Дипломник	_____	Дмитриевский А.С.

г. Долгопрудный, 2011

Содержание

Аннотация

В работе описывается постановка задачи, методология и текущее состояние проекта по созданию геополитической онтологии Российской Федерации — специального информационнопоискового тезауруса для автоматической обработки текстов с геополитической тематикой.

1 Введение

Информационные системы (ИС) разрабатываемые в последнее время характеризуются возрастающей сложностью устройства. Достаточно развитые ИС зачастую имеют широкую область охвата и предполагают использование онтологий или хотя бы глоссариев, тезаурусов или словарей ключевых терминов.

Россия — крупнейшее государство на планете и в официальных документах (конституции РФ, федеральных законах и пр.) используются сотни геополитических терминов, которые имеют субординацию и состоят в многочисленных отношениях друг с другом. Таким образом государство само по себе может быть рассмотрено как некоторая очень развитая информационная система. В таком рассмотрении становится очевидной необходимость разработки онтологии данной системы. Любое государство в основе своей всегда имеет законы, обычно выступающие в роли постулатов, нормирующих все производные термины и отношения. Однако в нынешнем состоянии правовые акты РФ представляют в лучшем случае отформатированные и оцифрованные текстовые документы. Но это делает их малопригодными для автоматизированных поисковых систем, машин выводов и прочих систем, которым могут понадобиться данные, хранящиеся в таких документах.

1.1 Актуальность

В крупных проектах, как правило, принимает участие большое количество участников. И для создания корректных систем требуется, чтобы у всех участников проекта была единая терминология предметной области, так как даже словарные термины могут иметь разные определения в разных источниках. Так термин «подшипник» имеет неоднозначное определение согласно ГОСТ ИСО 4378-1-2001 и различные трактовки данного термина могут привести к тому, что подшипник будет выступать в роли опоры[1].

Чтобы избежать терминологических неточностей при создании сложных ИС, разработчики вынуждены составлять глоссарии, словари терминов, онтологии и иные концептуальные схемы, формализующие понятия, используемые для разработки и использования ИС.

В то же время объемы информации в сети интернет с каждым годом растут, хотя сама информация при этом остается малоструктурированной, так как представляет из себя простые текстовые документы. Так как современные поисковые системы не учитывают семантику запросов и индексируют документы, не учитывая семантику хранящихся ресурсов,

поиск по таким ресурсам бывает крайне неэффективен. Для решения задачи повышения эффективности поиска в сети интернет предлагается строить порталы знаний, каждый из которых предоставляет доступ к ресурсам сети определенной тематики. Основу таких порталов знаний составляют онтологии, содержащие описание структуры и типологии соответствующих сетевых ресурсов [2].

Государство — сложная система и эффективное управление им подразумевает четкую иерархию институтов управления. В России существуют одновременно несколько типов и уровней административного деления: субъекты Российской Федерации, федеральные округа, экономические районы военные округа и др., границы которых могут не совпадать. Это создает неоднозначности, затрудняющие классификацию терминов в автоматическом режиме. А ручная классификация становится невозможной при достаточно больших объемах информации или при наличии требования достаточно подробной классификации. Учитывая вышесказанное, разработка онтологии правовой и в частности геополитической системы Российской Федерации является актуальной задачей.

1.2 Цель

Целью работы является разработка непротиворечивой онтологии геополитической системы Российской Федерации. Онтология должна быть максимально полной, чтобы быть приспособленной для использования автоматизированными обработчиками новостных потоков и текстов и содержала в себе как термины (напр. «населенный пункт»), так и сущности («Калужская область»).

Попутно в данной работе будет проведен краткий обзор систем создания и языков представления онтологий, позволяющих проводить максимально эффективную разработку онтологий.

1.3 Научная новизна

В настоящее время существуют исследования и разработка в области создания геополитических онтологий. В частности, Продовольственная и сельскохозяйственная организация Объединенных Наций разработала общемировую геополитическую онтологию `http://www.fao.org/countryprofiles/geoinfo.asp`. Однако, данная онтология охватывает весь земной шар и относится скорее к международной геополитике, а геополитическое деление отдельных государств в ней не представлено. Иван Бегтин, занимающийся созданием электронного правительства, также уделял внимание данной онтологии и добился некоторых результатов,

часть которых мы будем использовать в нашей работе. Но в целом до сих пор в России результаты данных работ используются слабо, хотя и имеют довольно большой потенциал. В США уже предпринимались попытки по построению и внедрению геополитических онтологии. Самый крупный проект в этой области на данный момент - проект OEGOV `http://www.oegov.us/`. Поэтому конкретно для России данная работа является довольно новым проектом.

1.4 Практическая ценность

Разработка такой онтологии будет полезна для автоматической обработки и индексации множества текстов связанных с ГС РФ, которая в свою очередь нужна для структуризации и классификации новостей и правовых актов в автоматическом режиме. Корректное решение данной задачи могло бы значительно улучшить результаты автоматической категоризации текстов. Результативность поисковых систем может быть значительно повышена именно в результате корректного анализа текстов и данных находящихся в сети. Помимо этого решение данной задачи способствовало бы развитию федеральной целевой программы «Электронная Россия», так как геополитическая терминология является неотъемлемой частью любого аспекта управлении государством. Данная работа может представлять интерес также для исследований в области неогеографии, так как неогеография предполагает современные методы представления геоданных и онтологические связи между географическими терминами могут быть эффективно интегрированы в неогеографические ИС

В качестве средств тематического поиска информации в течение многих лет использовались информационнопоисковые тезарусы. Но такие тезаурусы создавались для их использования в ручном индексировании и поиске, и не обеспечивают эффективный поиск в автоматическом режиме[3]. Так что использование онтологий - актуальная на сегодняшний день задача. К тому же, благодаря наличию связей и взаимоотношений между терминами в онтологиях, появляется возможность подключения машин вывода к онтологическим базам. Машины вывода позволяют производить первичную обработку информации на основе заложенных в онтологию аксиом, выявляя противоречия и делая заключения. Поэтому реализация геополитической онтологии Российской Федерации является ценным и практически нужным проектом.

2 Теоретическая справка

Для начала введем определение термина онтология согласно [4]:

Определение 2.1 *Онтология — эксплицитная спецификация концептуализации.*

Формально онтология состоит из терминов, организованных в таксономию, их определении и атрибутов, а также связанных с ними аксиом и правил вывода [5]. Формальная модель онтологии

$$O = \langle X, R, F \rangle$$

это упорядоченная тройка множеств, где:

- X — конечное множество понятий предметной области,
- R — конечное множество отношений между понятиями,
- F — конечное множество функций интерпретации.

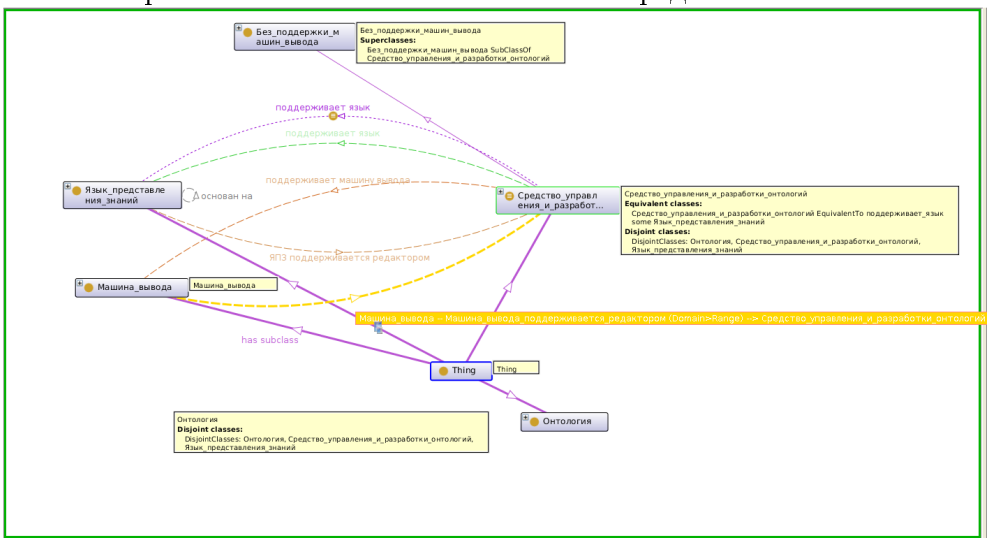
Множество X, по сути, словарь терминов используемых в предметной области (пример термина: «Средство управления и разработки онтологий» на рисунке 1), которую описывает онтология. R - множество отношений, которые связывают термины онтологии друг между другом (пример отношения: «ЯПЗ поддерживается редактором» на рисунке 1). А функция интерпретации F почти никогда в онтологии явно не задается.

Для наглядного представления основных составляющих онтологии, в процессе разработки была разработана онтология по средствам разработки онтологии и языкам представления онтологий. Некоторая часть онтологии представлена на рисунке 1.

Классы в этой онтологии обозначены прямоугольниками с желтыми кружками. Корневой элемент онтологии — класс «Thing». Любая существующая сущность или понятие является вещью (принадлежит классу «Thing»). Тем самым можно считать, что принадлежность термина классу «Thing» постулирует факт принадлежности данного термина области компетенции данной онтологии.

Любой класс связан стрелками с другими классами онтологии. Стрелки представляют отношения между классами. Любой класс, кроме класса «Thing» связан отношениями, хотя бы с родительским классом посредством отношения «has subclass»-«has superclass». На рисунке 1 можно увидеть несколько типов отношений помимо «has subclass» («has superclass» на рисунке не указан так как является обратным отношением к «has subclass») таких как:

Рис. 1: Простейшая онтология по языкам представления онтологий



- поддерживает язык,
- поддерживается редактором,
- основан на,
- поддерживает машину вывода;

Более простыми словами:

Определение 2.2 *Онтология — формальное представление знаний, как множества терминов в пределах определенной предметной области, а также связей установленных между терминами.*

Очевидно, что чем шире терминологическая база какой-либо области, тем сложнее будет установить различные связи между терминами и в особенности управлять данной ИС. Поэтому для эффективного управления знаниями необходим высокий уровень структурированности и организации данных. В этой ситуации использование онтологий оказываются одним из наиболее эффективных решений. Именно на основе использования онтологических терминов планируется сделать «машинную обработку смысла» контента максимально четкой, например, в рамках проекта Semantic Web. Таким образом, онтологии рассматриваются как основополагающая технология для использования в Semantic Web [6].

2.1 Языки описания онтологий

Для реализации различных онтологии, необходимо разработать языки их представления с достаточной выразительной мощностью, позволяющие избежать «низкоуровневых» проблем при проектировке онтологии. Цель таких языков — дать возможность указывать дополнительную машинно-интерпретируемую семантику ресурсов, сделать машинное представление данных более похожим на положение вещей в реальном мире, существенно повысить выразительные возможности концептуального моделирования слабо структурированных Web-данных [6].

Существуют традиционные языки спецификации онтологий:

- Ontolingua (`http://www.ksl.stanford.edu/software/ontolingua/`),
- CycL (`http://www.cyc.com/cycdoc/ref/cycl-syntax.html`),
- LOOM (`http://www.isi.edu/isd/LOOM/`),
- OKBC (`http://www.ai.sri.com/~okbc/`),
- OCML (`http://technologies.kmi.open.ac.uk/ocml/`),
- FLORA-2 (`http://flora.sourceforge.net/`),
- Более поздние языки, основаные на Web-стандартах:
  - XOL [7],
  - SHOE [8],
  - UPML [9];
- Специально для обмена онтологиями через Web:
  - RDF(S) [10],
  - DAML (`http://www.daml.org/about.html`),
  - OIL (`http://xml.coverpages.org/oil.html`),
  - OWL [11];

В целом, различие между традиционными и Web-языками спецификации онтологии заключается в выразительных возможностях описания предметной области и некоторых возможностях механизма логического вывода для этих языков. Типичные примитивы языков дополнительно включают:

- конструкции для агрегирования, множественных иерархий классов, правил вывода, аксиом;
- различные формы модуляризации для записи онтологий и взаимоотношений между ними;
- возможность мета-описания онтологии, что полезно при установлении отношений между различными видами онтологий.

2.1.1 Язык RDF

В рамках проекта семантической интерпретации информационных ресурсов Интернет (Semantic Web) был предложен стандарт описания метаданных о документе Resource Description Framework, использующий XML-синтаксис.

В RDF используется базовая модель данных «объект—атрибут—значение». позволяющая ему сыграть роль универсального языка описания семантики и связей для различных ресурсов. Ресурсы описываются в виде ориентированного размеченного графа — каждый ресурс может иметь свойства, которые в свою очередь также могут быть ресурсами или их коллекциями. Все словари RDF используют базовую структуру, описывающую классы ресурсов и типы связей между ними. Это позволяет использовать разнородные децентрализованные словари, созданные для машинной обработки по разным принципам и методам. Важной особенностью стандарта является расширяемость: можно задать структуру описания источника, используя и расширяя такие встроенные понятия RDF-схем, как классы, свойства, типы, коллекции. Модель схемы RDF включает наследование классов и свойств.

RDF Schema — стандарт, предложенный по инициативе W3C для представления онтологических знаний. Он специфицирует множество всевозможных допустимых схем данных. Модели предметных областей описываются посредством ресурсов, свойств и их значений. RDFS предоставляет хорошие базовые возможности для описания словарей типов предметных областей. Одно из ограничений — невозможность с помощью RDFS выразить аксиоматические знания, т. е. задать аксиомы и правила вывода, построенные на них.

2.1.2 Язык DAML+OIL

DAML+OIL — семантический язык разметки Web-ресурсов, расширяющий стандарты RDF и RDF Schema за счет более мощных примитивов моделирования. Последняя версия DAML+OIL обеспечивает бога-

тый набор конструкций для создания онтологии и разметки информации таким образом, чтобы их могла читать и понимать машина.

Первыми предложениями по описанию онтологии на базе RDFS были DARPA DAML-ONT (DARPA Agent Markup Language) и European Commission OIL (Ontology Inference Layer). Эти стандарты спецификации и обмена онтологиями были разработаны для поддержки процесса обмена знаниями и интеграции знаний. На базе этих предложений и возникло совместное решение DAML+OIL. Онтология DAML+OIL состоит из:

- заголовков (headers);
- элементов классов (class elements);
- элементов свойств (property elements);
- экземпляров (instances).

2.1.3 Язык OWL (Web Ontology Language)

OWL — язык представления онтологий, расширяющий возможности XML, RDF, RDF Schema и DAML+OIL. Этот проект предусматривает создание мощного механизма семантического анализа. Планируется, что в нем будут устранены ограничения конструкций DAML+OIL. Язык Веб-Онтологий OWL разработан для использования приложениями, которые должны обрабатывать содержимое информации, а не только представлять эту информацию людям. OWL обеспечивает более полную машинную обработку Веб-контента, чем та, которую поддерживают XML, RDF, и RDF Schema (RDF-S), предоставляя наряду с формальной семантикой дополнительный терминологический словарь. OWL имеет три диалекта (в порядке возрастания выразительности): OWL Lite, OWL DL и OWL Full [12].

OWL Lite использует только некоторые из особенностей языка OWL и имеет больше ограничений на использование этих особенностей, чем OWL DL или OWL Full. Например, в OWL Lite классы могут только быть определены в терминах именованных суперклассов (суперклассы не могут быть произвольными выражениями), и могут использоваться только определенные виды ограничений класса. Эквивалентность между классами и соподчинение между классами также разрешается только между именованными классами, а не между произвольными выражениями класса. Точно так же, ограничения в OWL Lite используют только именованные классы. OWL Lite также имеет ограниченное понятие кардинальности - разрешены только кардинальности 0 или 1.

OWL DL и OWL Full используют один и тот же словарь, хотя OWL DL подчинен некоторым ограничениям. Грубо говоря, OWL DL требует разделения типа (класс не может также быть индивидом или свойством, а свойство не может быть также индивидом или классом). Это подразумевает, что ограничения не могут налагаться на сами языковые элементы OWL (то, что позволено в OWL Full). Кроме того, OWL DL требует, чтобы свойства были определены или как Свойства-Объекты, или как Свойства-Значения: Свойства-значения — это отношения между представителями классов и литералами RDF и типами данных XML Schema, в то время как Свойства-объекты - это отношения между представителями двух классов.

Онтологии OWL — это последовательности аксиом и фактов, а также ссылок на другие онтологии. Они содержат компоненту для записи авторства и другой подробной информации, являются документами Web, на них можно ссылаться через URI.

2.2 Современные системы разработки онтологий

В данной работе мы будем использовать язык описания онтологий OWL. Для эффективной разработки онтологий на данный момент создано множество систем разработки, редакторов и пр. В процессе решения нашей задачи было опробовано более 5 систем разработки онтологий, серди которых

- Protégé (`http://protege.stanford.edu/`)
- NeOn Toolkit (`http://neon-toolkit.org/wiki/Main_Page`)
- TopBraid Composer (`http://www.topquadrant.com/products/TB_Composer.html`)
- OBO-Edit2 (`http://www.oboedit.org/`)

Для обзора особенностей этих и некоторых других систем разработки онтологии была составлена специальная онтология, отображающая концептуальную схему онтологических языков и редакторов. Исходный код данной онтологии представлен в приложении 5.B.

Рассмотрим отдельные системы разработки

2.2.1 Protégé

Наиболее популярным средством онтологического моделирования и создания баз знаний является свободно распространяемый редактор онтологий Protégé разработки Стенфордского университета. Разработка

Protégé исторически определялась направленностью на приложения в биологии и медицине, так как само приложение было разработано в отделении Медицинской Информатики Школы Медицины Стэнфордского Университета. Но при это система является независимой от предметной области и может быть использована в любых областях. Protégé позволяет создавать и редактировать файлы в поддерживаемых форматах RDF, RDFS, OWL, правила SWRL и SPARQL запросы с помощью удобного графического интерфейса и позволяет генерировать html-документы, отображающие структуру онтологий. Кроме того, рассматриваемое средство может использоваться как среда для выполнения правил, запросов, механизмов рассуждений. С помощью доступных механизмов рассуждений они позволяют выполнять классификацию и проверку логической целостности на основе OWL. Данная система разработки онтологий имеет поддержку нескольких машин вывода (HermiT, FacT++, Pellet, OWLlink HTTPXML и др.), также очень гибкая система пользовательского интерфейса позволяет делать настройки интерфейса очень специфичными под любую задачу. Допускается выполнение SPARQL запросов как над RDF хранилищами, так и над интегрированными реляционными базами данных. Поскольку он использует фреймовую модель представления знаний OKBC, это позволяет адаптировать его и для редактирования моделей предметных областей, представленных не в OWL, а в других форматах (UML, XML, SHOE, DAML+OIL, RDF и RDFS и т. п.) [13]. Таким образом, в Protégé применяется один настраиваемый интерфейс для обработки языков семантической разметки. Одним из таких языков является OWL. Это делает Protégé привлекательной средой разработки OWL описаний онтологий, тем более что при необходимости Protégé позволяет перевести все существующие наработки в Protege на другие языки семантической разметки, например, RDF. Кроме того, форма записи RDF-конструкций, генерируемых средой Protégé, является более предсказуемой, в отличие от разнообразных форм записи одних и тех же RDF-конструкций, сокращенных форм, которые можно построить с помощью обычного текстового редактора, что делает среду Protégé привлекательной для разработчиков решений, связанных именно с обработкой онтологий. Предлагаемая технология обработки RDF графов ориентирована в большей степени на обработку графовых конструкций Protégé-реализации стандарта RDF [14, 15]. Однако это обилие возможностей несколько загромождает систему, делая иногда слишком долгим процесс перенастройки. Тем не менее из всех опробованных систем данная оказалась наиболее удобной для реализации поставленной задачи. Самое большое преимущество данной системы заключено в ее модульности и большом количестве плагинов, написанных под данную

систему и расширяющих его возможности.

2.2.2 NeOn Toolkit

Данная система создания и редактирования онтологий находится в процессе доработки. На данный момент NeOn Toolkit разрабатывается параллельно с проектом `http://www.neon-project.org`. Данное средство имеет более четкую структуру пользовательского интерфейса и создания онтологий, однако навигация и редактирование объектов сделаны неудобно изза отсутствия возможности перехода по ссылкам. Поддерживает системы вывода HermiT и Pellet.

2.2.3 TopBraid Composer

Данный редактор является коммерческим, некоммерческая версия вряд ли может рассматриваться как эффективное средство разработки онтологий, ввиду отсутствия поддержки визуализации, машин вывода и др. Для выполнения правил и вывода дополнительных отношений между ресурсами в TopBraid Composer используется внутренний движок правил Jena. Однако, платная же версия поддерживает также машины вывода SWRL and Jena rules, TopSpin, Swift OWLIM, Pellet. Однако, данная система, предоставляя большие возможности в управлении онтологиями, не позволяет достаточно эффективно редактировать их.

2.2.4 OBO-Edit2

Данная система имеет очень удобный интерфейс для разработки и редактирования онтологий на языке OBO, в то же время данная система является неудобной для разработки на языке OWL вследствие отсутствия в бесплатной версии естественных для языка таких возможностей, как переход по ссылке на объект или экземпляр, автодополнение наименований класса и др.

2.3 Обзор аналогичных работ

Любая задача возникает вследствие определенных причин и по мере развития проблемы предпринимаются попытки разрешения ее. Так в США уже предпринималась попытка создания геополитической онтологии. На момент написания данной работы онтологии проекта eGovernment располагались на официальном веб сайте по адресу `http://www.oegov.us/`. Однако, специализация под геополитику США и акцентирование

внимания в первую очередь на социальных институтах фактически свели на нет применимость онтологии к локализации геополитических терминов. Так, например, иерархия административных единиц в этой онтологии состоит из 5 сущностей (Страна, Регион, Дистрикт, Штат и Провинция), наша же онтология ставит целью более общий охват терминов. Геополитическая онтология США в формате OWL/RDF доступна по адресу `http://www.oegov.org/core/owl/oe2gov`.

Имеются и другие примеры аналогичных работ в данной области. Так, например, Продовольственная и сельскохозяйственная организация Объединенных Наций (`www.fao.org`) обнародовала окончательную геополитическую онтологию, доступную и на русском языке. Онтология имеет географическое и политическое разделение стран, коды ООН и ISO 3166 для каждой страны, макроэкономические и другие показатели о странах и их регионах. Онтология доступна на официальном веб сайте Продовольственной и сельскохозяйственной организации ООН `http://aims.fao.org/aos/geopolitical.owl`. Однако, чрезмерный охват онтологии, не позволяет применять ее в масштабах одной страны со спецификой терминов и таксономией административных единиц каждого отдельного государства, Потому такая онтология может быть полезна как онтология более высокого уровня и использоваться при объединении локальных онтологий со всемирной онтологией геополитических терминов. Так для онтологии geopolitical минимальная структурная единица это «геополитический регион».

Онтологии И. Бегтина так же преследовали целью создание геополитической онтологии, однако они не являются непротиворечивыми, что приводит к тому, что машины вывода не могут работать на данных онтологиях. И следовательно мы их можем использовать лишь как источник первичной информации, но данные онтологии непригодны для слияния.

3 Построение геополитической онтологии РФ

3.1 Постановка задачи

Целью данной работы является разработка онтологии охватывающей все геополитические термины, используемые в Конституции РФ. Однако решение этой задачи требует дальнейшей работы вне данной работы. Сначала будет мы реализована упрощенная схематическая модель, позволяющая делать простейшие выводы и отслеживать противоречия. Непротиворечивость онтологии отслеживается главным образом с помощью машин выводов. В данной работе используются машины выводов

Pellet и HermiT. Геополитическая онтология РФ должна включать в себя административные единицы, субъекты федерации и отношения между ними, так чтобы можно было опираясь на данную онтологию при помощи машин вывода строить семантические карты например новостных потоков.

К сожалению, Российская Федерация не имеет единой системы административно-территориального деления, нормативного определения типологии населенных пунктов; в настоящее время здесь используются преимущественно советские критерии и принципы административно-территориального деления. Федеральное законодательство относит вопрос об административно-территориальном устройстве субъектов Российской Федерации к компетенции представительных (законодательных) органов субъектов Российской Федерации (подпункт «л» пункта 2 статьи 5 Федерального закона от 6 октября 1999г. (ред. от 9 февраля 2009г.) №184-ФЗ «Об общих принципах организации законодательных (представительных) органов субъектов Российской Федерации»; Определение Конституционного Суда Российской Федерации от 10 июля 2003г. №289-О). При этом для определения административно-территориального устройства субъектов Российской Федерации должны учитываться положения Федерального закона от 6 октября 2003г. №131-ФЗ «Об общих принципах организации местного самоуправления в Российской Федерации».

Вследствие этого при построении онтологии геополитики РФ, целесообразно строить ее на основе слияния локальных федеральных геополитических онтологий, так как в каждом субъекте может быть свое управление в зависимости от Констиутции субъекта или иных локальных правовых актов используемых в управлении субъектом.

3.2 Избыточные и неоднозначные определения в правовых актах РФ

Многочисленные правовые источники иногда вступают в конфликт друг с другом в отношении определения терминов. При построении онтологии необходимо разрешать подобного рода конфликты в пользу достаточной универсальности и гибкости онтологии, но в то же время не допуская превращения онтологии в онтологию высокого уровня. Главной проблемой при построении онтологии являются избыточность, неоднозначность и противоречивость входных данных терминологии.

3.2.1 Избыточность

Избыточность терминов на первый взгляд не представляет проблемы при создании ИС, так как избыточные, но непротиворечивые термины не могут нарушить логику системы. Однако, избыточность может привести к противоречию при коррекции правовых документов. В [16] написано

Определение 3.1 *Муниципальное образование — городское или сельское поселение, муниципальный район, городской округ либо внутригородская территория города федерального значения.*

Но в том же документе сказано:

Определение 3.2 *Городской округ — городское поселение, которое не входит в состав муниципального района и органы местного самоуправления которого осуществляют полномочия по решению установленных настоящим Федеральным законом вопросов местного значения поселения и вопросов местного значения муниципального района, а также могут осуществлять отдельные государственные полномочия, передаваемые органам местного самоуправления федеральными законами и законами субъектов Российской Федерации [16].*

Таким образом определение Муниципального образования избыточно, поскольку городской округ уже является городским поселением. Но в данном случае определение термина не вызывает противоречия.

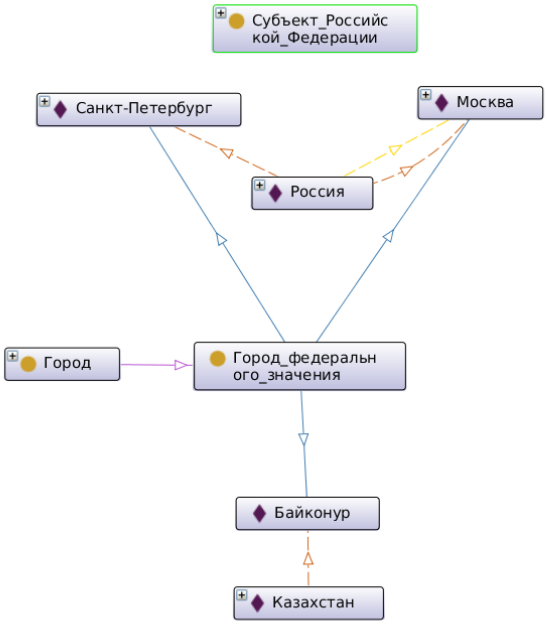
3.2.2 Противоречивость и неоднозначность

Теперь приведем пример противоречивой терминологии. Согласно юридическому словарю

Определение 3.3 *Город федерального значения — один из видов субъектов РФ...*

Рассмотрим город Байконур, расположенный в Казахстане, но арендованный Российской Федерацией до 2050 года. Согласно «Соглашению...» [17]: «На период аренды комплекса «Байконур» город Байконур в отношениях с Российской Федерацией наделяется статусом, соответствующим городу федерального значения Российской Федерации, с особым режимом безопасного функционирования объектов, предприятий и организаций, а также проживания граждан.» Таким образом мы приходим к противоречию: город Байконур является городом федерального значения РФ, но не является субъектом РФ (Рис. 2). Естественно для правовой

Рис. 2: Противоречие на примере термина «город федерального значения»



системы город Байконур в данном случае выступает как исключительный объект и не вызывает работы или сбоев в работе правовой системы. Так что для правовой системы Байконур является примером неоднозначной терминологии. Но с точки зрения формальной логики определение города Байконур оказывается некорректным и противоречивым, так что разумно избегать таких формулировок в правовых документах. Поэтому для разрешения конфликта мы были вынуждены не объявлять города федерального значения как подкласс субъектов РФ, хотя случай с городом Байконур скорее исключение из правила.

3.3 Описание исследования

После анализа современных языков представления онтологий для разработки нашей онтологии был выбран язык Ontology Web Language. Данный язык включает в себя возможности большинства его предшественников, и на данный момент OWL является языком описания онтологий рекомендуемых международным сообществом W3C для создания онтологий для веб. После анализа современных средств разработки

онтологий было решено вести разработку с использованием редактора Protégé. Данная система является одной из наиболее современных и в тоже время ее функционал достаточно удобен и широк для использования всех возможностей языка OWL.

Для построения таксономии терминов мы будем использовать в основном правовые документы РФ. Особое внимание при этом уделялось Конституции РФ (глава 1 статья 5) и федеральному закону «Об общих принципах организации местного самоуправления в Российской Федерации». Наибольшую проблему при построении данной онтологии представляет неорганизованность исходных данных, так как определения терминов разбросаны по множеству правовых актов и не определенны точно. Для примера можно привести подпункт 1 пункта 1 статьи 11 главы 2 Федерального закона Российской Федерации от 6 октября 2003 г. №131-ФЗ «Об общих принципах организации местного самоуправления в Российской Федерации»: «территория субъекта Российской Федерации, за исключением территорий с низкой плотностью населения, разграничивается между поселениями», что приводит к тому что некоторые поселения формально могут принадлежать разным субъектам федерации одновременно.

Важным моментом является разрешение неоднозначностей, система должна эффективно распознавать тип административного деления подразумеваемого по названию единицы. Эта задача относится к классификации текстов, решению данной задачи в общем виде посвящена работа [18].

Процесс построения онтологии распадается на серию подпроцессов. На первом этапе необходимо выделить основные классы сущностей и определить их значения на естественном языке. После этого на основе таксономических отношений (отношения «Класс-Подкласс») строятся деревья классификации понятий.

Для фиксации значимых отношений между терминами выделяются основные связи между ними, которые можно графически отобразить с помощью бинарных отношений. Помимо этого для геополитической онтологии интерес представляют отношения взаимного территориального положения (включающие в себя отношения «содержит_в_себе-входит_в_состав», «граничит_с» и пр.). Такие связи в дальнейшем могут послужить основой для интеграции различных онтологий.

Ну и наконец для практического использования нужны данные, характеризующие те или иные сущности онтологии. Данный этап мы планируем автоматизировать. На основании этих данных во многом определяется непротиворечивость онтологии.

И наконец, самым последним этапом в создании онтологии является

наполнение ее сущностями.

3.4 Таксономия административно-территориального деления в РФ

Основной целью геополитической онтологии РФ является установление таксономии терминов территориально-административного деления России и базисных отношений между терминами.

Хотя в Российской федерации административно-территориальное деление и определяется самостоятельно субъектами РФ, на первом этапе можно выделить в качестве административно-территориальных единиц

- Субъекты Российской Федерации
- Федеральные округа
- Экономические районы
- Муниципальное образование
- Военные округа

В качестве базисных отношений для территориальных единиц используются отношения:

- Входит_в_состав (is_member_of) – асимметричное свойство, с обратным свойством «содержит_в_себе»
- граничит с (border_on) – симметричное нерефлексивное свойство.

Данные отношения устанавливают наиболее общие взаимосвязи между сущностями онтологии, так как определяют географическое взаимоположение территориальных единиц и их состав.

3.5 Автоматизация процесса построения онтологии

Важным моментом при построении онтологии является эффективное использование уже существующих ресурсов. Так множество документов, в том числе и электронных (именно они и будут представлять для нас наибольший интерес), уже содержит массу нужной информации. Недостаток ее состоит в том, что информация неорганизованна или плохо структурирована. Таким образом для наполнения онтологии было принято решение использовать вспомогательные алгоритмы производящие структуризацию готовых источников. Информация о федеративном

устройстве Российской федерации была обработана через источники сети интернет. Для обработки были использованы алгоритмы написанные на языке PHP, с использованием библиотеки «PHP Simple HTML DOM Parser» для парсинга таблиц. Так из таблицы «Перечня субъектов федерации» страницы Википедии «Федеративное устройство России» была взята информация о типах федеральных субъектов, их именах, населении, площади территории, административных центрах и кодах ОКАТО. Идея парсинга заключалась в том, чтобы отформатировать имеющуюся информацию, структурированную в таблицах и на разрозненных страницах HTML, и перевести ее в XML/OWL-формат. По сути данная конвертация представляет собой structured retrieval — извлечение информации из структурированных документов (в XML формате) с акцентом на текстовую структуру документов [19].

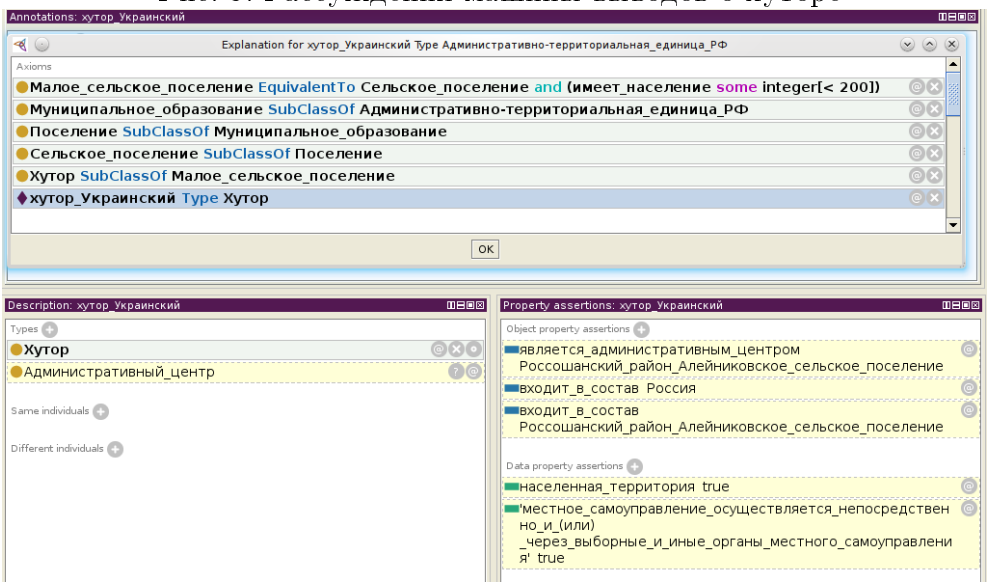
После этого машина вывода сможет дополнить онтологию на основании уже имеющихся данных и аксиом. Основные страницы проекта по парсингу данных приведены в приложении 5.А.

3.6 Пример использования онтологии

Для примера рассмотрим хутор Украинский, о котором известно, что он является хутором, административным центром Россошанского района Алейниковского сельского поселения. В соответствии с нынешним состоянием онтологии машина вывода на основе введенных данных и аксиом автоматически утверждает, что хутор Украинский является населенной территорией, имеет местное самоуправление осуществляющееся непосредственно или через выборные органы местного самоуправления, входит в состав Россашанского района. С помощью машины выводов Pellet (Incremental) было сделано заключение, что хутор Украинский является административно-территориальной единицей РФ. На рисунке 3 представлена последовательность заключений на основе которых система делает вывод о принадлежности хутора классу «Административно-территориальная единица РФ». Идея вывода полностью опирается на таксономию онтологии. Хутор в конечном счете всего лишь подкласс административно-территориальной единицы РФ.

Еще один более сложный пример автоматизированного вывода показывает возможные последствия некорректных аксиом в онтологиях. Изначально предполагалось, что каждая область имеет единственный административный центр. Это и было описано в качестве аксиомы в онтологии. Но после недолгой работы оказалось, что Московская область имеет органы государственной власти расположенные и в Москве и в Красногорске, причем Московская область не содержит в себе го-

Рис. 3: Рассуждения машины выводов о хуторе

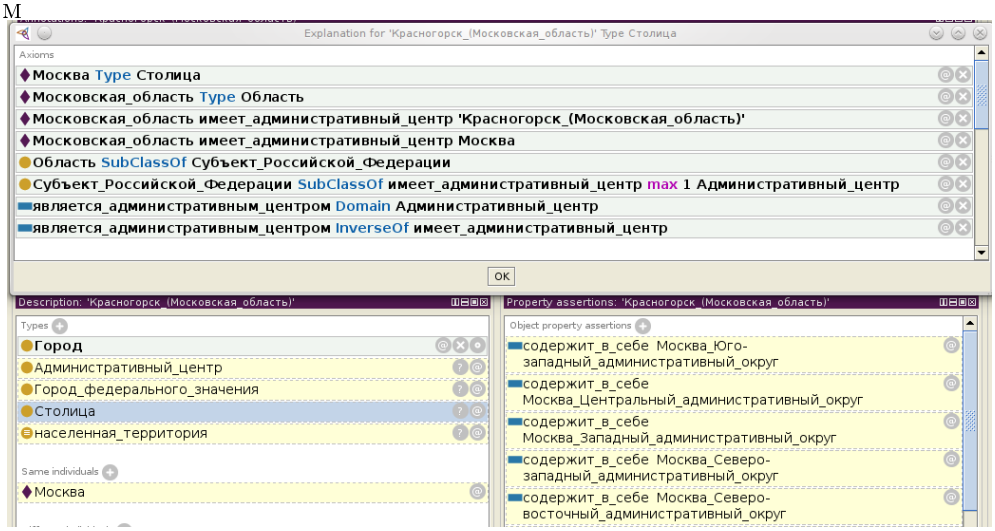


род Москва. Таким образом оказалось, что Московская область имеет два административных центра. Поскольку сначала база опиралась на старые аксиомы, то вскоре машина выводов сделала заключение, что экземпляр Красногорск — столица России. На рисунке 4 приведены подробные рассуждения машины вывода. В данном примере идея процесса вывода состоит в том, что, если город Красногорск и столица Москва — административный центр Московской области, а область может иметь лишь один административный центр, то города Москва и Красногорск — совпадающие экземпляры.

3.7 Анализ результатов

На данный момент составлена таксономия классов геополитических терминов РФ и набор важнейших межклассовых отношений. Начато наполнение терминов онтологии. Уже сейчас онтология в связке с машинами выводов (HermiT и Pellet) допускает получение некоторых выводов на основании неполных входных данных, опираясь исключительно на начальные аксиомы. Пример аксиом приведен на рисунке 5. Исходный код онтологии представлен в приложении 5.С.

Рис. 4: Ошибочные рассуждения машины выводов из-за неверных аксиом



4 Заключение

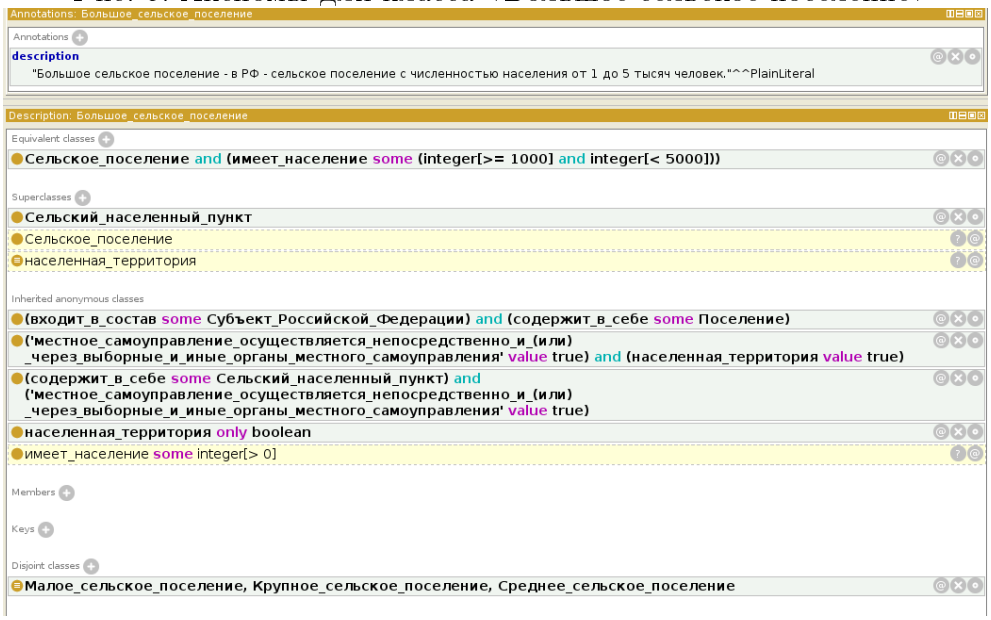
4.1 Выводы

В результате проделанной работы был изучены современные языки представления знаний (в частности OWL) и проведен сравнительный обзор современных инструментов по работе с онтологиями. Итоги данного исследования были занесены в специальную онтологию по языкам представления онтологий и средствам представления онтологий (см. 5.B). По мере исследования был проведен анализ предметной области и рассмотрены некоторые проблемы возникающие при разработке. Кроме этого был написан парсер для трансляции неструктурированного текста из некоторых веб-ресурсов в код OWL для автоматизированного наполнения онтологии данными. Данный парсер написан на языке PHP и приведен в приложении 5.A. На данный момент составлена достаточно детальная таксономия геополитических терминов (см. 5.C). Однако, для полной реализации поставленной цели требуются еще значительная работа по составлению связей между концепциями и завершение ввода терминов.

4.2 Перспективы

Результаты данной работы могут представлять интерес для автоматизированной обработки текстовой информации, использующей геополи-

Рис. 5: Аксиомы для класса «Большое сельское поселение»



тическую тематику. Под данное множество попадают как правовые акты относящиеся к РФ, так и новости. Эффективная автоматизированная обработка в свою очередь открывает перспективы для более совершенного анализа текстов в целом, классификации по тематикам новостей, поиску родственных новостей и многих других современных задач связанных с обработкой знаний. Так же данная работа, будучи тесно скоррелированной тематически с неогеографией России, может быть в дальнейшем использована для новейших карт России.

5 Приложения

5.A Парсер федеральных субъектов

Основной код программы парсинга данных о федеральных субъектов Российской Федерации.

```
1  <?php
2  include_once('simplehtmldom/simple_html_dom.php');
3
4  function handleNumber($str, $default) {
5      preg_match_all("/([0-9]|\.|\,)+/", $str, $matches);
6      $out = "";
7      foreach ($matches[0] as $stp) {
8          $out .= $stp;
9      }
10     sscanf($out, "%f", $num);
11     if (empty($num)) {
12         return $default;
13     } else {
14         return $num;
15     }
16 }
17 //
18
19 //
20 function fillFS($resp, $typeOfFS) {
21     include_once("FederalSubject.php");
22     $fs = new FederalSubject();
23     switch ($typeOfFS) {
24         case "Республики":
25             $fs->type = "Республика_в_составе_Российской_Федерации";
26             break;
27         case "Края":
28             $fs->type = "Край";
29             break;
30         case "Области":
31             $fs->type = "Область";
32             break;
33         case "Города федерального значения":
34             $fs->type = "Город_федерального_значения";
35             break;
```

```
36          case "Автономная область":
37              $fs->type = "Автономная_область";
38              break;
39          case "Автономные округа":
40              $fs->type = "Автономный_округ";
41              break;
42      }
43      $replacer="_";
44      $fs->name = str_replace(" ", $replacer, $resp->children(1)->children(0)->
45      $fs->territory = handleNumber($resp->children(4)->plaintext, 0);
46      $fs->population = handleNumber($resp->children(5)->plaintext, 0);
47      $fs->admCenter = str_replace(" ", $replacer, $resp->children(6)->children
48      $fs->OKATOcode = $resp->children(8)->plaintext;
49      return $fs;
50  }
51
52  function parseType($types, $type, $key, $typeOfFS) {
53
54  }
55
56  function parseTable($table) {
57      $types = $table->find('tr[style="background:#eee;"]');
58
59      $FSs = Array();
60      $allowedTypes = array("Республики", "Края", "Области", "Города федерально
61
62      if ($types != null) {
63          foreach ($types as $key => $type) {
64              $s = trim($type->plaintext);
65              foreach ($allowedTypes as $allowedType) {
66                  if (strstr($s, $allowedType)) {
67                      $resp = $type->nextSibling();
68                      while ($key + 1 < count($types) && $resp != $types[$key +
69                          $FSs[] = fillFS($resp, $allowedType);
70  //                        echo "name=" . $resp->children(1)->children(0)-> pl
71  //                        echo "terr=" . $resp->children(4) . "<br>";
72  //                        echo "popu=" . $resp->children(5) . "<br>";
73  //                        echo "admc=" . $resp->children(6)->children(0)->pla
74  //                        echo "okat=" . $resp->children(8) . "<br>";
75                          $resp = $resp->nextSibling();
76                      }
```

```
77                  }
78              }
79
80              //            $temp=$type->find('td[colspan="10"]');
81              //            if($temp!=null){
82
83              //            }
84          }
85          //        echo "bingo";
86          //        print_r($ns);
87      } else {
88          //        echo "NObingo";
89
90      }
91
92      $writer = new XMLWriter();
93      $writer->setIndent(true);
94      $writer->openMemory();
95      $writer->startDocument('1.0','UTF-8');
96  //    $str = '';
97
98      foreach ($FSs as $fs) {
99          $fs->convertToOWL($writer);
100     }
101     $writer->endDocument();
102     $xml = $writer->outputMemory();
103     if ($fedSubj = fopen('federalSubjects.owl', "w+")) {
104         fwrite($fedSubj, $xml);
105         fclose($fedSubj);
106     }
107 }
108
109
110
111 $html = file_get_html('http://ru.wikipedia.org/wiki/Федеративное_устройство_Р
112 $result = $html->find('table.sortable');
113
114 if (count($result) != 1) {
115     echo "Что-то изменилось на старнице <a href=http://ru.wikipedia.org/wiki/
116 } else {
117     $subHtml = new simple_html_dom();
```

```
118         $subHtml->load($result[0]->outertext);
119         include("FederalSubject.php");
120         $FSs = parseTable($subHtml);
121         echo "Grabbing Complete";
122     }
123
124
125
126     ?>
```

Метод **convertToOWL** представляет собой просто конвертер в OWL-формат.

Функция **parseTable** - парсер исходных данных.

5.B Онтология по онтологиям

Исходный код онтологии по онтологиям можно скачать по адресу: `http://sourceforge.net/projects/geopoliticalont/files/Ontology1291680129914.owl/download`

5.C Геополитическая онтология

Исходный код геополитической онтологии в формате OWL доступен по адресу: `http://sourceforge.net/projects/geopoliticalont/files/Ontology1292303228208.owl/download`

Список литературы

[1] М.Б. Кац. Что такое подшипник? *Конструктор-Машиностроитель*, 10:22–27, 2007.

[2] О.И. Боровикова and Ю.А. Загорулько. Организация порталов знаний на основе онтологий. *Компьютерная лингвистика и интеллектуальные технологии*, 2:76–82, 2002.

[3] Б.В. Добров and Н.В. Лукашевич. Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного-поиска. *Ученые записки казанского государственного университета*, 149(2):49–72, 2007.

[4] T. R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

[5] Т.А. Гаврилова and В.Ф. Хорошевский. *Базы знаний интеллектуальных систем*. СПб.: Питер, 2000.

[6] А.Я. Гладун and Ю.В. Рогушина. Онтологии в корпоративных системах. *Корпоративные системы*, №1:12, 2006.

[7] Vinay K. Chaudhri Peter D. Karp and Jerome Thomere. Xol: An xml-based ontology exchange language, August 31 1999.

[8] Jeff Heflin and Sean Luke. *Simple HTML Ontology Extension*. Parallel Understanding Systems Group Department of Computer Science University of Maryland at College Park, w3c edition.

[9] D. Fensel, M. Crubezy, F. Van Harmelen, and I. Horrocks. Oil & upml: A unifying framework for the knowledge web. In *In Proceedings of the Workshop on Applications of Ontologies and Problem-solving Methods, 14th European Conference on Artificial Intelligence ECAI'00*, pages 00–14, 2000.

[10] W3C. *RDF Vocabulary Description Language 1.0: RDF Schema*, 10 February 2004.

[11] W3C. *OWL 2 Web Ontology Language*, w3c owl working group edition, October 2009.

[12] Deborah L. McGuinness and Frank van Harmelen. *OWL Web Ontology Language Overview*. W3C, Февраль 2004.

[13] Б.В. Добров, В.В. Иванов, Н.В. Лукашевич, and В.Д. Соловьев. *Онтологии и тезаурусы: модели, инструменты, приложения.* Интуит.РУ, 2008.

[14] Н.В. Рябова and С.С. Щербак. Развитие технологий semantic web: обработка rdf-графов на основе xslt. *Щербак.Net*, 2008.

[15] Д.В. Левшин and А.С. Марков. Алгоритмы интеграции СУБД postgresql с семантическим веб. Московский Государственный Университет им. М.В. Ломоносова.

[16] *Федеральный закон об общих принципах организации местного самоуправления в Росийской Федерации*, октябрь 2003.

[17] *Соглашение между Российской Федерацией и Республикой Казахстан о статусе города Байконур, порядке формирования и статусе его органов исполнительной власти*, Москва, декабрь 1995.

[18] Alfio Gliozzo and Carlo Strapparava. *Semantic Domains in Computational Linguistics.* Springer-Verlag Berlin Heidelberg, 2009.

[19] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.