Онтологическое моделирование и публикация данных об Особо Охраняемых Природных Территориях

© К.А. Кузнецов © В.А. Серебряков © К.Б. Теймуразов
© Е.С. Устинова © Д. А. Малахов
Вычислительный центр им. А.А.Дородницына РАН,
г.Москва
K.Kuznetcov@gmail.com serebr@ccas.ru kbt@intring.ru
jane.echo90@gmail.com dimon-malakhov@yandex.ru

Аннотация

В статье рассматриваются проблемы публикации данных об Особо Охраняемых Природных Территориях (ООПТ) в пространстве Linked Open Data (LOD). Предлагается онтология данных об ООПТ, соответствующая отечественным и международным стандартам предметной области и удовлетворяющая рекомендациям проекта LOD. Также описана методика публикации и связывания данных об ООПТ с данными из внешних источников.

1 Введение

Интенсивный прогресс информационных технологий привел к тому, что все большие и большие объемы пространственных данных (т.е. данных о пространственных объектах, включающих сведения об их местоположении и свойствах) становятся доступными в сети Интернет. Можно отметить следующие важные особенности пространственных данных:

- Пространственные данные состоят из двух частей – географической информации, которая может быть представлена в векторном или растровом виде и снабжена метаданными, и непространственных атрибутов, которые определяют семантику пространственного объекта;
- Общие пространственные данные зачастую являются фактором, связывающим воедино данные из различных предметных областей;
- Геоинформационное сообщество выработало ряд стандартов на пространственные данные и метаданные. Большинство доступных в сети пространственных данных следуют им;
- В отличие от многих других предметных областей, в геоинформационном сообществе существует развитая культура свободного обмена данными, имеются многочисленные сервисы и инструменты обмена данными.

Благодаря перечисленным выше особенностям наборы пространственных данных достаточно легко переносятся из традиционного гипертекстового Веб в Семантический Веб. На практике это означает простоту включения геопространственных наборов данных в проект Linking Open Data [3], целью которого является наполнение сети Интернет данными в стандартных форматах Semantic Web [9], а также устанавливание связей между данными из различных источников. Таким образом формируется единое пространство данных Linked Open Data. Проект носит рекомендательный характер, описывает набор технологий и методик для работы с семантическими данными. Публикация данных в пространстве Linked Open Data позволяет увеличить степень повторного использования данных, понизить степень дублирования данных, повысить ценность данных за счет связывания их с другими данными и облегчить их потребление заинтересованными сторонами. По состоянию на осень 2011 года географические данные составляют примерно 20% от всех опубликованных в пространстве Linked Open Data данных, 10% наборов данных пространства Linked Open Data используют термины из словарей Basic Geo Vocabulary и GeoNames [8]. Таким образом, публикация наборов пространственных данных в пространстве Linked Open Data является актуальным направлением развития геоинформатики.

2 Постановка задачи

Работа посвящена конкретной тематической области геопространственных данных – особо охраняемым природным территориям (ООПТ). Целью работы является разработка прикладной схемы для описания пространственных данных ООПТ, построение демонстрационной базы геоданных на основе полученной схемы и создание приложения для публикации данных в пространстве Linked Open Data. Разрабатываемая прикладная схема должна:

- Удовлетворять требованиям федерального законодательства РФ, затрагивающим вопросы ООПТ [12];

Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.

- Быть совместимой с распространенными в мире стандартами на публикацию пространственных данных и данных об ООПТ в частности;
- Основываться на имеющихся в наличии наборов данных об ООПТ «Валдайский национальный парк» и «Национальный парк “Таганай”».

Для хранения данных необходимо разработать схему реляционной базы данных и ее отображение в прикладную схему, создать механизм загрузки данных в виде SHP-файлов в базу данных.

Для публикации данных в пространстве Linked Open Data необходимо, согласно рекомендациям проекта Linking Open Data:

- Описать прикладную схему на языке OWL [9] с использованием терминов распространенных в пространстве Linked Open Data словарей;
- Обеспечить механизм идентификации объектов ООПТ при помощи HTTP URI;
- Обеспечить механизм генерации RDF/XML документов с представлением объектов ООПТ;
- Обеспечить механизм дереференсирования этих URI, т.е. предоставления соответствующих RDF/XML документов в ответ на HTTP запросы;
- Установить и поддерживать связи со внешними ресурсами пространства Linked Open Data.

3 Модель данных

Первым этапом решения поставленной задачи является разработка онтологии ООПТ. Разработка осуществлялась следующим образом. Сначала были проанализированы имеющиеся наборы данных, отечественная нормативно-правовая база предметной области, а также международные стандарты, и была составлена концептуальная модель предметной области. Затем были проанализированы распространенные в пространстве Linked Open Data словари и наборы данных, релевантные предметной области, из них были отобраны термины, соответствующие концептуальной модели. На их основе была создана онтология, которая была дополнена новыми терминами для полного соответствия модели.

При анализе имеющихся в наличии наборов данных о национальных парках были выявлены четыре группы объектов - информация об объекте «охраняемая территория» (границы парка, охранной зоны, функциональное зонирование территории парка), информация об охраняемых объектах (места обитания охраняемых биологических видов), прочие нетематические объекты (железные дороги, газопроводы и т.п.) и вспомогательные объекты, связанные с тематическими.

Затем был проанализирован федеральный закон № 33-ФЗ «Об особо охраняемых природных территориях» от 14 марта 1995 г. [12] В результате были сформулированы требования к концептуальной модели:

- Должен быть создан класс «Особо охраняемая природная территория» («ООПТ»), имеющий обязательное свойство «граница», содержащее пространственные данные;
- Класс «ООПТ» должен иметь обязательный атрибут «категория» типа перечисление;
- Класс «ООПТ» должен иметь необязательное свойство «граница охранной зоны», содержащее пространственные данные;
- Класс «ООПТ» должен иметь обязательный атрибут «статус» типа перечисление;
- Должен быть создан класс «Функциональная Зона» с обязательным атрибутом «назначение зоны» типа перечисление и обязательным свойством «граница», содержащим пространственные данные;
- Класс «ООПТ» должен иметь необязательное свойство «выделеныеЗоны» неограниченной кардинальности типа «Функциональная Зона».

3.1 Используемые международные стандарты

В результате анализа международных стандартов выяснилось, что на настоящий момент в мире наиболее широко распространены следующие стандарты на публикацию пространственных данных: стандарты серии 19100 технического комитета ISO/TC 211 [10], стандарты OGC (Open Geospatial Consortium, Inc.), а также наборы стандартов CEN (Европа) и FGDC (США). Эти стандарты во многом похожи между собой, и практически полностью совместимы. Из международных стандартов на публикацию данных об ООПТ распространение получила только спецификация данных INSPIRE (Infrastructure for Spatial Information in Europe) [6]. Инициатива INSPIRE на законодательном уровне устанавливает стратегию развития общеевропейской инфраструктуры пространственных данных, а также стандартные наборы пространственных метаданных и правила взаимодействия пространственных сервисов. В числе определяемых стандартами INSPIRE тематических наборов метаданных есть и стандарт, смежный тематике ООПТ - INSPIRE Data Specifiaction on Protected Sites. Стандарт является последовательной, четко проработанной структурой, составленной ведущими европейскими специалистами по геоматике и согласуется с международной серией стандартов ISO 19100 – Geographic Information. Спецификации содержат формализованное описание модели предметной области в виде UML диаграмм классов и предполагают использование языка GML (Geography Markup Language) для кодирования данных. Спецификация INSPIRE хорошо совместима с принципами Linked Open Data - HTTP URI идентификаторы удовлетворяют требованиям INSPIRE, UML схемы INSPIRE и GML изоморфны RDF, принципы связывания ресурсов соответствуют определенным в стандарте INSPIRE Generic

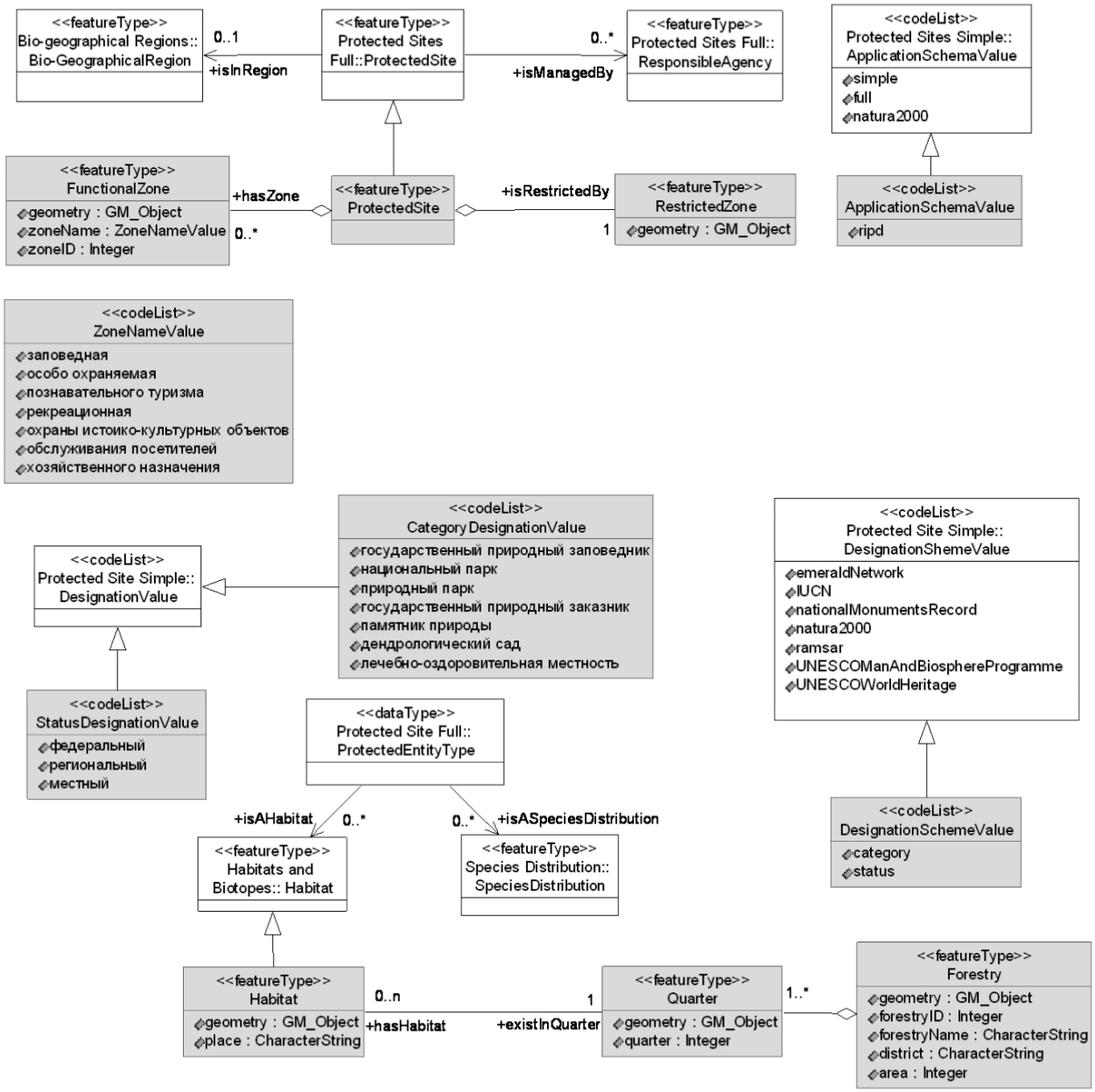

Рис. 1 Схема данных об ООПТ

Conceptual Model понятиям связей. Исходя из этого при разработке онтологии ООПТ было решено использовать стандарт INSPIRE Data Specifiaction on Protected Sites и связанные с ним стандарты Data Specification on Geographical Names (названия пространственных объектов) и Guidelines for the encoding of spatial data (кодирование пространственных данных). Спецификация данных INSPIRE по охраняемым территориям содержит 3 прикладных схемы, формализованные в виде UML диаграмм классов – «Simple», «Full» и «Natura 2000». В качестве основы для онтологии была выбрана схема «Full», как обеспечивающая максимально полный набор классов и атрибутов для описания предметной области. Было произведено сопоставление между элементами схемы данных INSPIRE, элементами имеющихся в наличии наборов данных, и требованиями, накладываемыми законодательством РФ. В результате из схемы INSPIRE «Full» была получена концептуальная модель данных об ООПТ, представленная на Рис.1 (серым цветом выделены добавленные элементы, атрибуты исходной схемы опущены).

Спецификация данных INSPIRE Data Specification on Geographical Names описывает понятия, связанные с географическими названиями, то есть именами собственными, применяющимися для обозначения существующих естественных, техногенных и культурных объектов. Спецификация определяет минимальное ядро, необходимое для описания названий пространственных объектов, и расширенную схему для наборов пространственных данных, несущих лингвистический характер. В случае концептуальной модели данных об ООПТ достаточно использовать ядро спецификации.

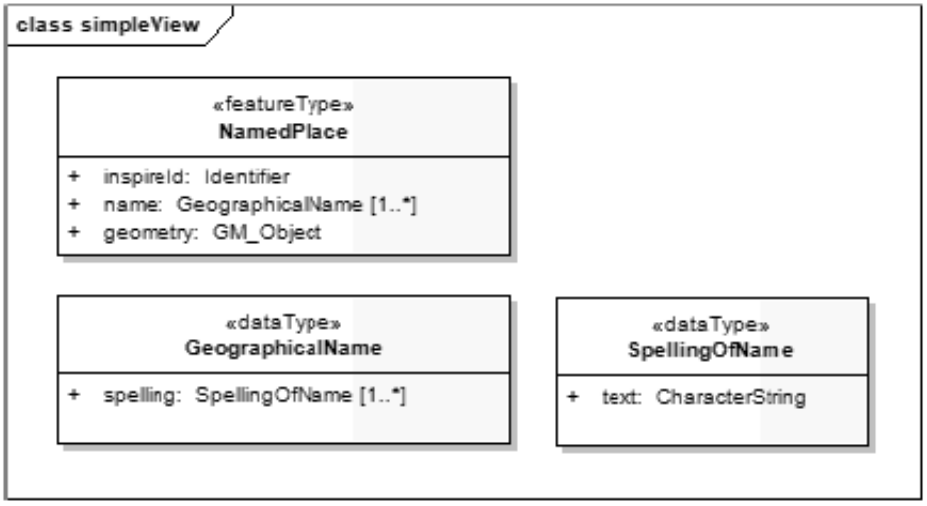

Рис. 2 Ядро прикладной схемы INSPIRE Data Specification on Geographical Names

Наконец, согласно спецификации INSPIRE «Guidelines for the encoding of spatial data», геометрия пространственных объектов должна быть представлена согласно стандарту ISO 19107, в котором определены различные геометрические объекты. Учитывая требования INSPIRE и специфику имеющихся данных в концептуальной модели данных должны быть представлены классы Point, Polygon, LinearRing, MultiPolygon. Класс Point должен быть описан двумя или тремя координатами. Класс LinearRing описывается последовательностью Point. Класс Polygon описывается несколькими LinearRing , один из которых - внешняя граница, другие образуют внутреннюю границу и должны быть внутри внешней границы. Класс MultiPolygon представляет из себя коллекцию не пересекающихся Polygon.

3.2 Онтология ООПТ в пространстве Linked Open Data

В результате анализа пространства Linked Open Data были найдены следующие словари и наборы данных, релевантные предметной области ООПТ:

- Набор данных о пространственных объектах и их названиях GeoNames [11] (в дальнейшем используется префикс geonames);
- Набор данных о биологических видах GeoSpecies [5] (префикс geospecies);
- RDF словарь W3C Basic Geo Vocabulary (префикс geo) [9];
- RDF словарь NeoGeo Geometry Ontology (префикс neogeo) [7].

В пространстве Linked Open Data самой распространенной онтологией для описания географических имен является онтология GeoNames, в ней определен класс geonames:Feature, совпадающий по семантике с классом NamedPlace спецификации INSPIRE Data Specification on Geographical Names, имеющий свойства geonames :alternateName типа string, описывающие варианты географических имен. Класс Feature при этом является потомком класса SpatialThing онтологии W3C Basic Geo. Однако в онтологии GeoNames отсутствуют аналоги классов GeographicalName и SpellingOfName спецификации INSPIRE, из этого следует, что класса Feature недостаточно для представления класса NamedPlace в онтологии ООПТ. Более подходящих онтологий для описания географических имен в пространстве Linked Open Data не выявлено. Поэтому в разрабатываемую онтологию добавлены классы GeographicalName и SpellingOfName, разработанные самостоятельно на основании спецификации INSPIRE.

Тем не менее, класс словарь Geonames широко используется в пространстве Linked Open Data для классификации пространственных объектов при помощи свойств geonames:featureClass и geonames:featureCode. Поэтому все классы, моделирующие пространственные объекты, относящиеся к области ООПТ (лесничество, охранная зона и т.п.) были унаследованы от geonames:Feature. Кроме того, добавлено ограничение на класс ProtectedSites онтологии, моделирующий Охряняемую Территорию, которое фиксирует его код в таксономии Geonames (L.RESW, “wildlife reserve”):

oopt:ProtectedSites subclassOf (geonames:featureClass value geonames:L)

oopt:ProtectedSites subclassOf (geonames:featureCode value geonames:L.RESW)

Для описания геометрии воспользуемся существующей онтологией NeoGeo. Словарь NeoGeo является результатом обсуждений относящихся с гео-данным и предназначен для унификации интеграции данных в области геометрии. В онтологии NeoGeo ней описаны классы Geometry, Polygon, LineString, LinearRing, MultiPolygon, BoundingBox. Онтология использует класс Point, определенный в онтологии W3C Basic Geo. В W3C Basic Geo подразумевается система координат WGS84. Так как данные об ООПТ не обязательно ограничиваются этой системой координат, необходимо добавить необязательное

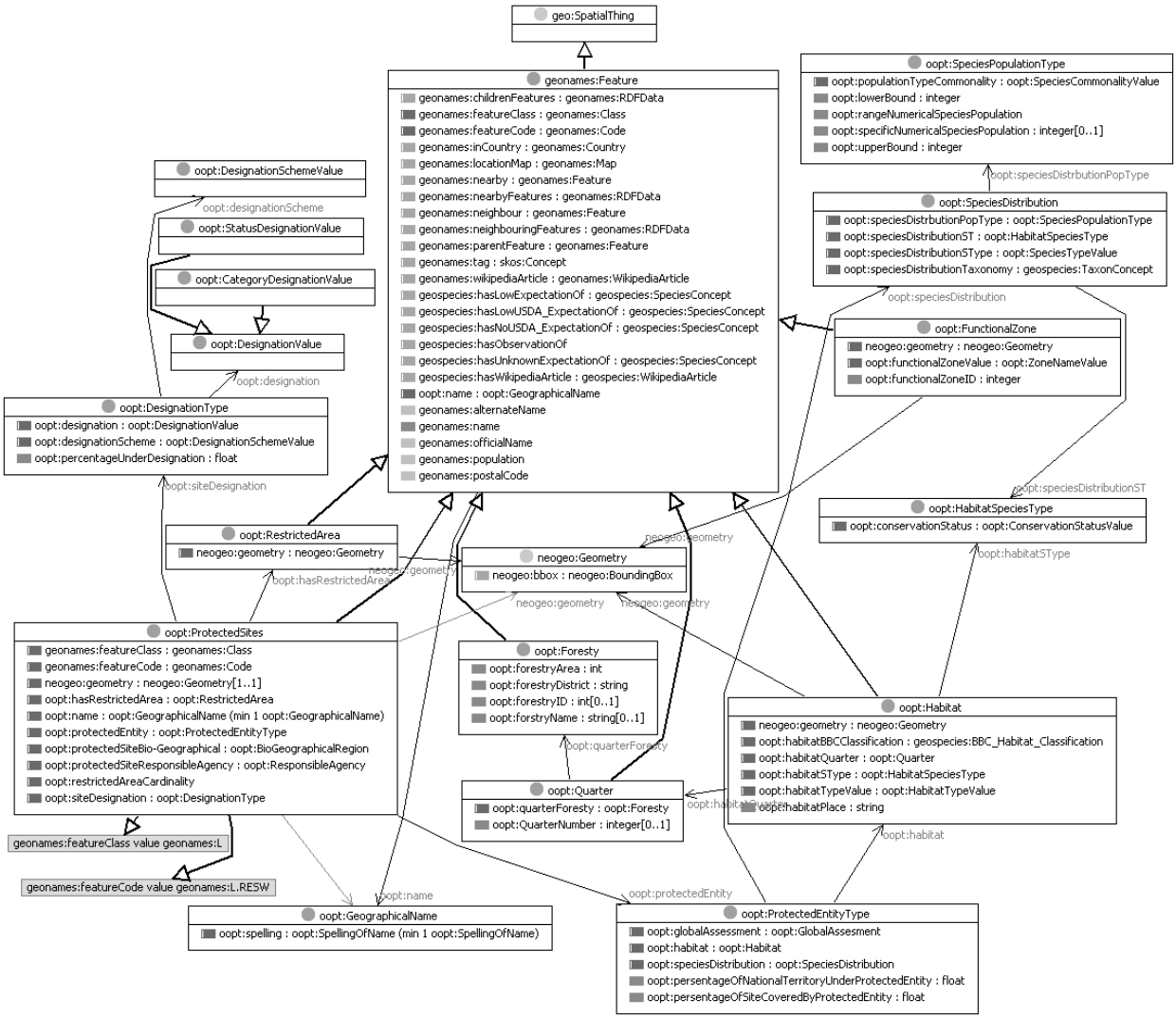

Рис. 3 Онтология ООПТ

свойство SC_CRS к классу Point, являющееся идентификатором системы координат, описанным в стандарте ISO 19111. В случае, если значение свойства не указано, считается что система координат WGS84. В стандарте ISO 19107 это поле определено у всех объектов геометрии, но в нашей ситуации это избыточность данных и возможность неоднозначности данных. Например, если Polygon выражается через несколько классов Point, которые имеют разные системы координат. Поэтому мы ограничиваемся классом Point. Так же необходимо добавить свойство point в класс LineString, которое показывает, что точка принадлежит ломаной. Свойство должно быть помечено как обратное для свойства partOf класса Point.

Онтология GeoSpecies используется для классификации ареалов обитания животных (при помощи классификатора geospecies:BBC_Habitat_Classification) и видов животных (при помощи классификатора geospecies:TaxonConcept). Остальные классы и свойства онтологии были созданы самостоятельно на основе спецификации INSPIRE. Диаграмма классов и свойств онтологии представлена на Рис.3 (некоторые классы-классификаторы опущены).

4 Публикация и связывание данных

После того, как онтология разработана, необходимо опубликовать данные в пространстве Linked Open Data в терминах полученной онтологии. Для этого исходные данные из SHP-файлов были загружены в реляционную базу данных с поддержкой пространственных типов данных (PostGIS). Схема реляционной базы данных была разработана вручную на основе онтологии и структуры SHP-файлов. Для публикации данных из реляционной базы данных в RDF\XML виде был выбран D2R Server [2], как наиболее простое некоммерческое решение для публикации RDF документов, поддерживающие дереференсирование HTTP URI ресурсов. Файлы отображения, необходимые для работы D2R Server были сгенерированы им автоматически и затем доработаны вручную для совместимости с онтологией.

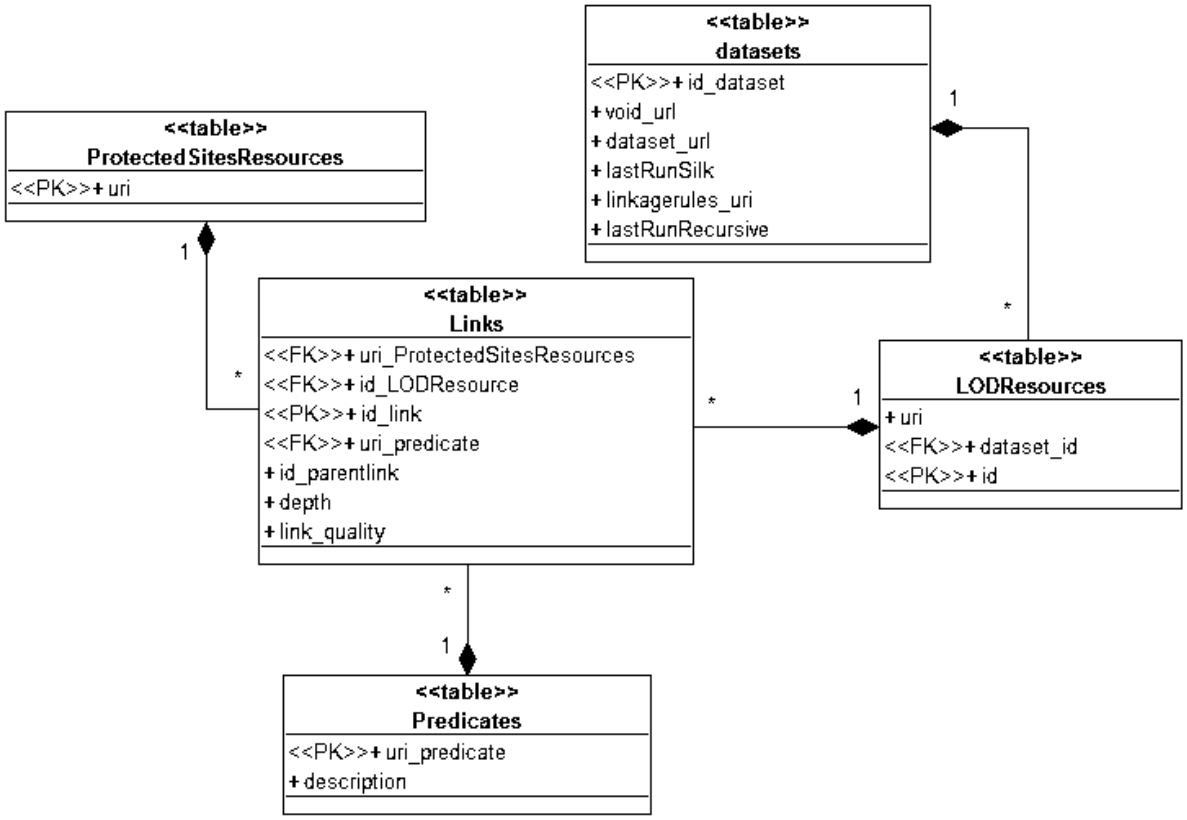

Рис. 4 Схема хранилища связей

4.1 Связывание данных

Согласно рекомендациям проекта Linking Open Data опубликованные данные должны быть связаны с данными из других наборов пространства Linked Open Data. В терминах RDF это означает, что RDF представления объектов из набора данных об ООПТ содержат RDF-тройки, субъектом которых является ресурс из этого набора данных, а объектом – ресурс из стороннего набора данных. Предикат же тройки определяет тип связи. В работе рассматриваются только предикаты, осуществляющие связь идентичных объектов - owl:sameAs и skos:closeMatch. Заметим, что проблема использования owl:sameAs для установления связей в пространстве Linked Open Data здесь не рассматривается.

Было разработано прототипное приложение, которое в полуавтоматическом режиме генерирует связи набора данных об ООПТ со внешними наборами данных. Приложение работает следующим образом. На первом этапе определяется целевой набор данных для поиска связей, и для этого набора вручную создается файл конфигурации связывания на языке SILK LSL [4], который содержит сведения о доступе к внешнему набору данных и правила связывания, состоящие из путей к сравниваемым свойствам в rdf-документе и метрикам, по которым оценивается близость значений заданных свойств. Приложение запускается вручную, и при помощи Silk Link Discovery Framework генерирует прямые связи между набором данных об ООПТ и заданным внешним набором данных. Затем приложение траверсирует RDF представления внешних ресурсов, находя связи owl:sameAs, rdf:seeAlso и skos:closeMatch с ранее связанными ресурсами (т.е. связи следующих порядков). Связи группируются по внешним наборам данных, которые идентифицируются при помощи их VoID-дескрипторов [1]. Глубина поиска задается в конфигурации. Найденные связи сохраняются в реляционной базе данных, откуда затем публикуются при помощи D2R Server.

Внешние наборы данных для поиска связей попадают в две основные категории – наборы данных о растениях/животных и наборы пространственных данных. Связи с наборами из первой категории устанавливаются достаточно просто на основании латинских наименований видов. В качестве исходных наборов для генерации связей используются Geospecies, DBPedia, Bio2RDF. Связи с наборами пространственных данных могут быть установлены либо по наименованиям объектов, либо по координатам ресурсов. Наименования пространственных объектов ООПТ включают наименования собственно охраняемый территорий («Валдайский национальный парк») и наименования ареалов обитания (названия рек, озер и т.п.). Для установления связей по координатам используются пространственные меры идентичности языка SILK LSL. Для генераций связей первого уровня по координатам

используются наборы данных Geonames и Linked Geo Data.

Перечисленные выше наборы данных являются центрами кластеров данных своих предметных областей в пространстве Linked Open Data. Связав набор данных об ООПТ с этими наборами данных мы получаем цепочки связей, ведущий от данных об ООПТ к данным из других предметных областей. Однако эти наборы данных не содержат исходящих связей с другими прикладными наборами данных об ООПТ, поэтому установить дальнейшие связи невозможно. На настоящий момент тематических наборов данных об ООПТ в пространстве Linked Open Data не существует, поэтому генерация исходящих связей с выбранными наборами позволяет достичь приемлемого уровня связанности со внешними ресурсами. Однако разработки в этом направлении ведутся, и в скором времени следует ожидать появления в пространстве Linked Open Data различных наборов данных об ООПТ.

5 Заключение

Результатом проделанной работы является онтология данных об ООПТ, соответствующая требования законодательства РФ и удовлетворяющая стандартам INSPIRE, а также простейшая система публикации и связывания данных в пространстве Linked Open Data, использующая разработанную онтологию.

Разработанная онтология может быть взята за основу при публикации в пространстве Linked Open Data пространственных данных из других прикладных областей, в особенности, тех, которые попадают под различные спецификации INSPIRE. Система публикации и связывания может быть адаптирована для любой предметной области.

Направления дальнейших работ включают:

- Улучшение механизма публикации данных, так, например, данные могут быть дополнены voiD-дескрипторами, а набор данных зарегистрирован в каталоге CKAN;
- Разработка пользовательского веб-интерфейса для просмотра и загрузки данных;
- Исследование вопроса автоматической генерации онтологии, схемы реляционной базы данных и правил отображения между ними по прикладной GML схеме;
- Улучшение механизма генерации связей;
- Добавление возможности интеграции и совместной публикации в пространстве Linked Open Data данных из различных независимых источников данных об ООПТ.

Литература

[1] Alexander K., Cyganiak R., Hausenblas M., Zhao J. Describing linked datasets. In Proceedings of the WWW2009 Workshop on Linked Data on the Web, 2009.

[2] Bizer C., Cyganiak R. D2R Server - Publishing Relational Databases on the Semantic Web. Poster at the 5th International Semantic Web Conference, Athens, USA, 2006. http://www4.wiwiss.fu-berlin.de/bizer/pub/Bizer-Cyganiak-D2R-Server-ISWC2006.pdf

[3] Heath T., Bizer C. Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool, 2011. http://linkeddatabook.com/editions/1.0/

[4] Volz J., Bizer C., Gaedke M., Kobilarov G. Discovering and maintaining links on the web of data. In Proceedings of the International Semantic Web Conference, pages 650–665, 2009.

[5] Сайт GeoSpecies Knowledge Base: http://about.geospecies.org/

[6] Сайт INSPIRE - Infrastructure for Spatial Information in Europe http://inspire.jrc.ec.europa.eu/

[7] Сайт NeoGeo Geometry Ontology: http://geovocab.org/geometry.html

[8] Сайт State of the LOD Cloud: http://www4.wiwiss.fu-berlin.de/lodcloud/state/

[9] Сайт World Wide Web Consortium (W3C): http://www.w3.org/TR/owl-features/

[10] Сайт коммитета ISO/TC 211 http://www.isotc211.org/

[11] Сайт проекта GeoNames: http://www.geonames.org/

[12] Федеральный закон. № 33-ФЗ «Об особо охраняемых природных территориях» от 14 марта 1995 г (по состоянию на 01.01.2010) http://www.legis.ru/misc/doc/312/

Modeling Ontology and Publishing Data on Protected Sites

K. Kuznetcov, V. Serebriakov, K. Teymurazov, E. Ustinova, D. Malakhov

The paper deals with problems of publishing of data on Protected Sites in Linked Open Data space. We introduce an OWL ontology for data on Protected Sites, which follows legislative system of Russian Federation and Linking Open Data project recommendations. The ontology uses common RDF vocabulary terms and adapts INSPIRE data model. We also present system for publishing this data and interlinking it with data from other RDF data sources.