

Публикация данных об Особо Охраняемых Природных Территориях в пространстве Linked Open Data

К.А. Кузнецов, В.А.Серебряков, К.Б.Теймуразов

Аннотация

В этой статье предлагается проект системы публикации данных об Особо Охраняемых Природных Территориях (ООПТ) в пространстве Linked Open Data. Описана общая архитектура системы, принципы работы модулей связывания, публикации и интеграции данных. Также предлагается онтология ООПТ, основанная на европейских стандартах INSPIRE.

Ключевые слова: система интеграции пространственных данные, Linked Open Data, наборы RDF-связей, подсистема публикации данных, связывание данных.

1. Введение

1.1. Пространство Linked Open Data

Задача интеграции данных заключается в соединении данных из различных источников и предоставлении пользователям унифицированного представления этих данных, в том числе возможности извлечения интересующей пользователя информации по запросу. Сложность задачи обуславливается автономностью и гетерогенностью источников данных. Можно выделить три последовательных уровня гетерогенности – физический (форматы файлов, кодировки и т.п.), логический (структурные различия между моделями данных) и семантический (различия в интерпретации элементов данных). С развитием информационных технологий проблемы семантической неоднородности становятся все более актуальными. В контексте глобальной сети Интернет решение задач интеграции данных привело к появлению понятия пространства данных (Web of Data) – расширению традиционного гипертекстового веб («пространства документов»), в котором информация представляется в строго формализованном виде, пригодном для машинной обработки. Таким образом решается проблема семантической неоднородности данных.

Практическим воплощением концепции Web Of Data является проект Linking Open Data (LOD) [12]. Целью этого проекта является наполнение сети Интернет данными в стандартных форматах Semantic Web, а также устанавливание связей между данными из различных источников. Таким образом формируется единое пространство данных Linked Open Data. Проект носит рекомендательный характер, описывает стек технологий и методик для работы с семантическими данными.

Стек технологий LOD включает в себя следующие стандарты W3C: протокол передачи данных HTTP, механизм глобальной уникальной идентификации URI, язык представления моделей данных в форме объектов и их взаимосвязей RDF [21], его расширения RDFS и язык веб-онтологий OWL [20], а также SPARQL [23] - протокол доступа и язык поисковых запросов к источникам данных RDF. В проекте рекомендуются четыре принципа функционирования пространства данных (Linked Data Principles):

- Необходимо использовать URI для идентификации не только веб-документов, но и объектов реального мира и абстрактных понятий.
- Необходимо использовать HTTP URIs для доступа к информации о ресурсах.
- В ответ на запрос HTTP URI необходимо предоставлять информацию о ресурсе в RDF формате.
- В этой информации необходимо предоставлять RDF-ссылки на другие ресурсы.

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 статистическим данным на сентябрь 2011 года [24], пространство Linked Open Data включает около 300 больших источников данных, 31*10⁹ RDF-троек, 500*10⁶ RDF-связей. Основные категории данных – научные данные, правительственные данные и контент, генерируемый пользователями.

1.2. Пространственные данные и Linked Open Data

Развитие геоинформатики привело к тому, что все большие и большие объемы пространственных данных становятся доступными в сети Интернет. Под пространственными данными здесь понимаются данные о пространственных объектах, состоящие 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, поэтому они играют важную роль в формировании глобального пространства данных. Геоинформационное сообщество выработало и широко использует ряд стандартов на моделирование и публикацию пространственных данных (стандарты OGC, ISO TC 211, INSPIRE), что практически снимает вопросы физической и логической гетерогенности данных. Для преодоления семантической гетерогенности был введено понятие Инфраструктуры Пространственных Данных (ИПД), были разработаны стандарты для построения распределенных систем на основе сервисов. Хотя это и облегчило совместное использование данных, большинство ИПД до сих пор не связаны друг с другом, а также с источниками непространственных данных.

Таким образом, публикация наборов пространственных данных в пространстве Linked Open Data является актуальным направлением развития геоинформатики. На текущий момент данные наук о Земле составляют около 10% всех источников

данных пространства Linked Open Data, около 20% от общего числа RDF-троек и 7% от общего числа RDF-связей. 10% наборов данных пространства Linked Open Data используют термины из словарей Basic Geo Vocabulary и GeoNames. Однако большинство публикуемых в пространстве Linked Open Data данных слабо соотносятся с выработанными стандартами ИПД.

2. Постановка задачи

Учитывая вышеизложенное, предлагается создать публичный ресурс пространственных данных, интегрированный в пространство LOD посредством автоматизированной системы. В качестве предметной области выбраны данные об Особо Охраняемых Природных Территориях (ООПТ). Разрабатываемая система будет предназначена для:

- формирования и поддержки схемы данных об ООПТ с учетом требований законодательства РФ, международных и отечественных стандартов на пространственные данные;
- хранения набора данных об ООПТ и оперативного обновления этого набора данных из различных «внутренних» источников, а также загрузка новых данных;
- публикации сформированного набора данных в сети Интернет в формате RDF с возможностью доступа к данным сформированного набора (включая связанные с ними данные других наборов) через интерфейс пользователя и интерфейс прикладных программ (API);
- предоставления стандартизированного доступа к данным при помощи WMS и WFS сервисов;
- предоставления единого интерфейса семантического поиска ко всем источникам данных системы.

Под «внутренними» источниками данных понимаются источники структурированных данных, которые не включены в пространство Linked Open Data и предоставляют API для поисковых запросов в терминах своей схемы данных. Система должна поддерживать различные типы источников данных, должна быть предусмотрена возможность добавления новых типов источников данных, а также подключения новых источников данных в процессе работы системы. Должны поддерживаться реляционные хранилища данных с возможностью JDBC доступа, SPARQL-точки доступа и WFS сервисы. Система должна поддерживать импорт данных из ESRI Shape-файлов, эффективно хранить и обрабатывать геометрические данные.

В качестве модели данных система должна использовать онтологию предметной области ООПТ, разработанную согласно рекомендациям проекта LOD. Онтология должна удовлетворять отечественным и международным стандартам на данные предметной области.

3. Обзор существующих решений

На настоящий момент не существует системы, которая обеспечивала бы одновременно всю рекомендуемую проектом LOD функциональность. Наиболее

комплексным решением является платформа Virtuoso Universal Server [27]. Она включает средства для представления информации из различных источников (реляционных баз данных, RDF-хранилищ, источников с Web API) в форме виртуальной базы данных, с возможностью публикации данных в RDF формате. Virtuoso поддерживает SPARQL доступ к данным и содержит такие компоненты, как RDF-краулер и простейший движок логического вывода. Virtuoso может дополнять публикуемые RDF-ресурсы VoID дескрипторами [1] и информацией о словарях данных. Однако Virtuoso является коммерческим продуктом, а его open-source версия обладает значительно урезанными возможностями. Среди open-source приложений для публикации данных в пространстве Linked Data следует упомянуть D2R Server [4] – инструмент для публикации RDF данных из реляционной базы данных с поддержкой SPARQL доступа. Также упоминается проект MASTRO [3], позволяющий предоставить SPARQL-доступ к реляционной базе данных. По сравнению с Virtuoso и D2R Server MASTRO предоставляет более выразительные средства для описания отображений между OWL-онтологией и реляционной базой данных, однако он не предоставляет никаких средств для публикации и связывания данных. Существуют также простейшие приложения для публикации RDF-данных из SPARQL-точек доступа и хранилищ RDF-троек (Pubby, Paget и т.п.).

Однако ни Virtuoso, ни D2R Server не предоставляют средств для установления и поддержки связей между RDF-ресурсами, как внутренними, так и внешними, ограничиваясь генерацией URI по настраиваемым шаблонам. Использование таких жестких шаблонов не всегда позволяет выявить отношения идентичности между элементами данных из различных источников и в полной мере представить их в виде связанных наборов данных Linked Open Data. Существует ряд независимых приложений для генерации и поддержки связей между элементами данных, таких как SILK [28], LINES [17], SemMF и DSNotify. Однако на настоящий момент не существует приложений, которые бы осуществляли интеграцию набора данных в пространство Linked Open Data, т.е. автоматически обнаруживали бы новые наборы данных и по возможности устанавливали и поддерживали связи с элементами данных из этих наборов данных. Возможность построения такой системы обсуждается в [18] и [26].

Возможности нетривиального практического применения сгенерированных наборов RDF-связей исследованы слабо. Можно отметить систему SPLENDID [10], которая использует статистику наборов связей для исполнения федеративных SPARQL запросов. Также наборы связей (Linksets в терминах VoID) применяются различными семантическими поисковыми системами.

4. Архитектура системы

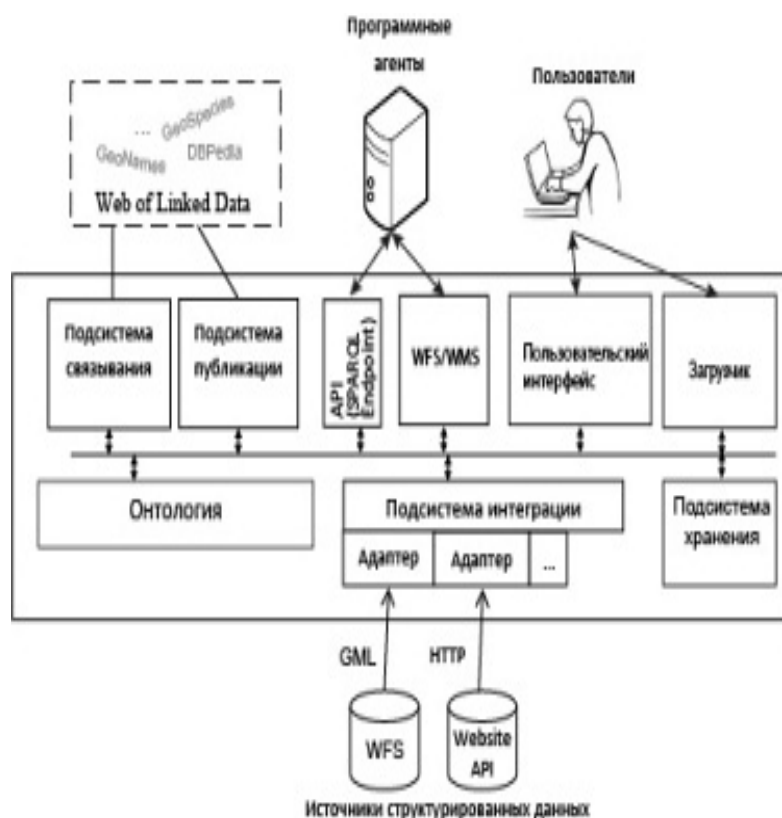
Для решения поставленной задачи предлагается разработать модульную систему, которая будет включать следующие компоненты:

- онтологию информационных объектов и связей, которые могут представлять интерес для пользователей и потребителей информации системы;
- подсистему связывания, предназначенную для связывания данных из

источников системы с данными из открытых наборов данных, опубликованных в пространстве Linked Open Data в формате RDF, а также между собой;

- подсистему публикации, предназначенную для предоставления пользователям и приложениям доступа к ресурсам из набора данных системы согласно рекомендациям проекта LOD;
- подсистему интеграции данных, предназначенную для предоставления единого интерфейса доступа ко всем источникам данных системы, а также для идентификации ресурсов системы;
- набор адаптеров для предоставления унифицированного SPARQL доступа к источникам данных различных типов (реляционные базы данных, Web API, WFS сервисов и т.п.);
- подсистему хранения, предназначенную для интеграции собранной информации и сгенерированных ссылок, организации семантического хранилища знаний в формате RDF, хранения системных настроек и служебной информации;
- загрузчик, позволяющий добавлять в систему данные из файлов;
- WFS и WMS сервисы, для предоставления доступа к данным без использования технологий Linked Open Data;
- веб-портал для доступа пользователей и API (SPARQL точку доступа) для программных агентов.

Архитектура системы представлена на Рис.1.



5. Структуры данных

5.1. Концептуальная модель данных

Первым шагом разработки системы является создание концептуальной модели данных. Для этого были проанализированы имеющиеся примеры наборов данных об ООПТ, отечественная нормативно-правовая база предметной области, а также международные стандарты. 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. [31] В результате были уточнены и дополнены наборы обязательных атрибутов выявленных объектов.

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, 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). В числе определяемых стандартами INSPIRE тематических наборов метаданных есть и стандарт, смежный тематике ООПТ - INSPIRE Data Specification on Protected Sites [14]. 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, что позволяет достаточно просто разработать по ним онтологию. Исходя из этого при разработке онтологии ООПТ было решено использовать стандарт INSPIRE Data Specification on Protected Sites и связанные с ним стандарты Data Specification on Geographical Names [13] (названия пространственных объектов) и Guidelines for the encoding of spatial data (кодирование пространственных данных) [15], Habitats and Biotopes (ареалы обитания и биотопы), Species Distribution (распределение видов), Biogeographical Regions (био-географические регионы).

Схема INSPIRE «Full» была дополнена ранее определенными классами и свойствами, в результате была получена концептуальная модель данных об ООПТ, представленная на Рис.2 (серым цветом выделены добавленные элементы, атрибуты исходной схемы опущены).

из схемы данных SHP-файла в схему реляционной БД, либо же преобразовывать SHP-файл в GML и уже его загружать в базу. При публикации RDF используются правила отображения из реляционной БД в RDF. Т.к. схема базы данных и онтология являются выражением концептуальной модели при помощи языков DDL и OWL, то теоретически возможна генерация всех схем данных и правил отображения между ними на основе прикладной схемы GML. Однако на практике существуют препятствия, которые делают автоматическую генерацию невозможной или нецелесообразной.

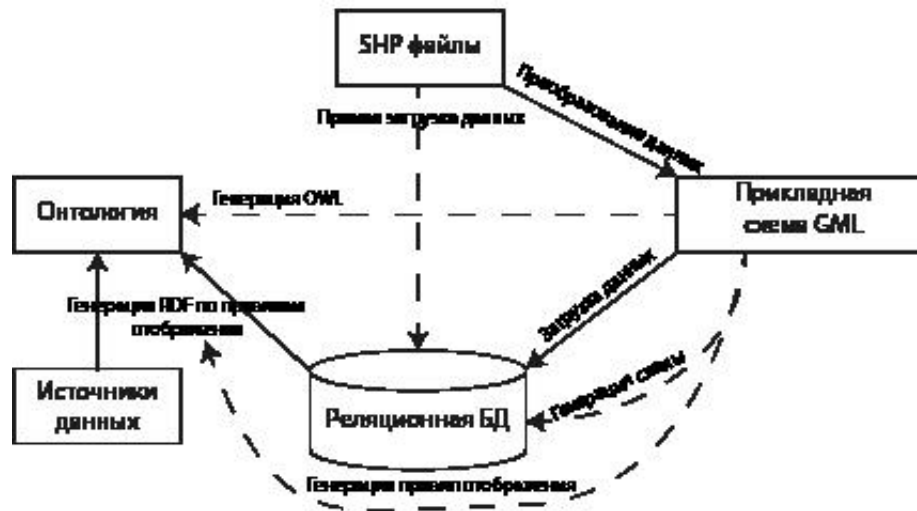


Рис. 3 Схемы данных в системе. Сплошные линии – выбранные решения, пунктирные – возможные пути реализации.

5.3. Онтология

RDF и GML во многом изоморфны, первые спецификации GML содержали в себе RDF/XML форму записи. Однако в последней версии спецификации такая форма записи не предусмотрена, и не существует стандартизированной онтологии на основе GML. Вопросы генерации онтологий по прикладным схемам GML рассмотрены в [25], однако предложенный в этой статье подход не учитывает специфику онтологий пространства Linked Open Data, а именно необходимость использовать широко распространенные словари. Из-за этого автоматическая генерация OWL невозможна, необходимо вручную подбирать словари и термины для моделирования предметной области.

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 [7] (в дальнейшем используется префикс geonames);
- Набор данных о биологических видах GeoSpecies [9] (префикс geospecies);
- RDF словарь W3C Basic Geo Vocabulary [29] (префикс geo);
- RDF словарь NeoGeo Geometry Ontology [16] (префикс neogeo);
- Стандарт OGC – словарь GeoSPARQL Vocabulary [8].

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, охранный зона и т.п.) были унаследованы от `geonames:Feature`. Кроме того, добавлено ограничение на класс `ProtectedSites` онтологии, моделирующий Охранный 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 к гео-данным и предназначен для унификации интеграции данных в области геометрии. В онтологии NeoGeo 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 свойство `SC_CRS` к классу `Point`, являющееся идентификатором системы координат, описанным в стандарте ISO 19111. В случае, если значение свойства не указано, считается что используется 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`. Необходимо также добавить свойство `point` в класс `LineString`, которое показывает, что точка принадлежит ломаной. Свойство должно быть помечено как обратное для свойства `partOf` класса `Point`. Стоит также отметить словарь недавно опубликованного стандарта OGC GeoSPARQL, в котором содержатся классы `SpatialObject`, `Feature` и `Geometry`. Этот словарь в целом идентичен словарю NeoGeo, однако не распространен в пространстве LinkedData.

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 прикладной схемы GML.

5.4. Схема базы данных и загрузчик

Существуют инструменты для генерации схемы реляционной базы данных по произвольной XML-схеме (XML Spy, GO Loader). Однако для сложных прикладных схем GML полученная структура базы является неэффективной, в случае схемы ООПТ полученная схема будет содержать множество ненужных элементов. Поэтому схема базы данных создана вручную на основе концептуальной модели.

Рассмотрим теперь процесс загрузки SHP-файлов. С концептуальной точки зрения SHP-файл - это набор однородных географических объектов. Объект представлен геометрической информацией и набором атрибутов. Поддерживаемые геометрические типы - точка, полилиния, полигон. Отдельно взятый SHP-файл не может содержать каких либо объектных связей, но SHP-файлы являются форматом экспорта из более сложных хранилищ геоданных, которые определяют такие отношения. При экспорте пакета SHP-файлов связи перестают быть явно определенными и сохраняются в виде атрибутов - внешних ключей. Для сохранения связей необходима пакетная загрузка SHP-файлов с ручным указанием связей.

Существует взаимно-однозначное отображение между GML-схемой и SHP-файлом. Существуют программные средства (например, GDAL со своим расширением OGR [19]), позволяющие из SHP-файла получить его GML-схему в формате XSD и GML-файл с данными. Далее возможны две альтернативы при работе загрузчика - трансформация данных в прикладную схему GML ООПТ с последующей загрузкой в базу данных, либо же прямая загрузка GML-файлов в базу данных системы. Для трансформации можно использовать XSLT-преобразования, но она в любом случае не может быть произведена полностью автоматически, т.к. затрагивается семантика данных. Однако, учитывая одноуровневый характер SHP-файлов, можно вручную произвести сопоставление классов объектов (Features) файлов с классами прикладной GML-схемы, указав по необходимости литеральные значения и внешние ключи. Использование прикладной GML схемы в качестве промежуточного этапа загрузки позволяет сделать моделирование более интуитивно понятным и освобождает от необходимости знания схемы базы данных. Кроме того это более универсальное решение, на основе которого можно в дальнейшем реализовать загрузку данных из других форматов. Поэтому при реализации загрузчика выбран первый вариант.

6. Публикация данных

Подсистема публикации является входной точкой в систему для пользователей и программных агентов Linked Data. Задача подсистемы публикации заключается в деривировании URI ресурсов системы, т.е. в предоставлении информации об этих ресурсах в ответ на HTTP запросы их URI. URI ресурсов в системе носят неинформационный характер, т.е. URI ресурса не связан напрямую с каким-либо его цифровым представлением. При обращении к таким URI подсистема

публикации данных осуществляет процесс, известный как content negotiation, т.е. перенаправляет клиента на URI нужного ему представления ресурса. В зависимости от заголовка HTTP GET запроса подсистема возвращает либо HTML-представление ресурса для обычных гипертекстовых браузеров, либо RDF/XML представление для приложений Linked Data (например, семантических браузеров). В первом случае к запросу будет добавлен заголовок Accept: text/html, во втором - Accept: application/rdf+xml.

Подсистема публикации данных получает от подсистемы интеграции данных информацию о ресурсе и затем генерирует на основе этих данных HTML+RDFa или RDF/XML представление ресурса. У подсистемы интеграции данных будет запрошена следующая информация:

- Все литеральные свойства ресурса, входящие и исходящие RDF-ссылки на этот ресурс. Эта информация может быть получена в результате простых SPARQL запросов с паттернами { ?x ?y } и { ?x ?y }. По URI ресурса система интеграции данных определит релевантный источник данных и перенаправит SPARQL запрос ему, а также при помощи набора связей определит идентичные ресурсы из других наборов и направит аналогичные запросы в релевантные источники.
- Класс ресурса и все вышестоящие по онтологии классы. Эта информация не всегда может быть получена при помощи SPARQL-запросов к подсистеме интеграции, поэтому онтология должна быть опубликована в явном виде (достаточно одного статического RDF-файла)
- Информация о наборе данных, к которому относится запрашиваемый ресурс.

В случае RDF/XML документа вся полученная информация объединяется в один RDF файл и возвращается клиенту. В случае HTML документа информация форматируется на основе шаблона, содержащего также RDFa атрибуты. Шаблоны в подсистеме публикации могут задаваться как для каждого класса в отдельности, так и в общем виде. Помимо публикации данных в пространстве Linked Open Data, система должна поддерживать и WMS/WFS сервисы, для этого можно воспользоваться платформой GeoServer или аналогичным инструментом.

7. Интеграция данных

Задача подсистемы интеграции – предоставить другим подсистемам или же внешним программным агентам единый интерфейс доступа к информации, хранимой в источниках данных системы. Требуемая информация специфицируется при помощи SPARQL запроса. Эта подсистема ответственна также за представление данных системы в виде единого набора данных пространства данных Linked Open Data. Существует несколько подходов к построению систем интеграции данных – хранилища данных, системы на базе посредников, децентрализованные системы. Поскольку система ориентирована на работу с многочисленными, сильно автономными источниками данных, был выбран подход интеграции данных по принципу посредников. Недостатком подобных систем (например, Virtuoso) является большой объем сетевых взаимодействий, возникающий в процессе ответа на поисковые запросы. Особенностью системы

является использование принципов Linked Open Data в такой архитектуре, что способно сгладить этот недостаток.

7.1. Наборы данных

Источники данных подключаются к системе при помощи компонентов-адаптеров, которые являются SPARQL-точками доступа к источникам данных и исполняют запросы в терминах системной онтологии. Адаптеры являются типовыми, настраиваемыми компонентами (например, JDBC адаптер, WFS адаптер). В отличие от существующих систем федерации данных (например Virtuoso с его Sponger картриджами), данные из различных источников не представляются единым набором данных с единым пространством идентификаторов.

При подключении и настройке адаптеров отношения идентичности между элементами данных из разных источников не устанавливаются за счет общих уникальных идентификаторов, а в духе Linked Open Data считается, что каждый источник содержит уникальные ресурсы. Каждому источнику данных выделяется подпространство имен системы - `http:///datasets/`, адаптер сопоставляет каждому ресурсу источника данных HTTP URI из этого подпространства имен. Таким образом, по URI ресурса системы можно установить его происхождение. Поскольку онтология системы имеет таксономический характер, не содержит сложных конструкторов концептов и не предполагает логического вывода, возможно также в URI ресурса явно отразить его класс. В общем виде HTTP URI ресурса выглядит следующим образом:

`http:///datasets///`

Такая схема URI согласуется с правилами идентификации ресурсов INSPIRE ID.

Помимо типовых настроек адаптера при подключении нового источника данных в систему указываются его общее описание (в терминах Dublin Core [5]), тематическая классификация, информация об авторских правах и т.п. В конфигурацию адаптера входит также информация о классах и свойствах системной онтологии, относящихся к источнику данных. Эта информация может быть введена вручную либо получена автоматически при помощи SPARQL ASK запросов. Свойства адаптера затем публикуются в виде VoID дескриптора набора данных и делаются доступными через подсистему публикации данных. Таким образом, каждый источник данных системы снаружи виден как отдельный набор данных Linked Data, однако все наборы данных имеют общую точку доступа через систему. При этом все данные системы представляются в виде одного набора данных, а наборы данных источников являются его потомками (в терминах VoID). Такая структура позволяет отразить автономность и независимость источников при их включении в пространство LOD.

8. Исполнение поисковых запросов

Исполнение поисковых запросов подсистемой интеграции осуществляется следующим образом [30]. На первом шаге производится переформулировка SPARQL запроса относительно аксиом глобальной онтологии. Построение такого

переформулированного запроса является своего рода логическим выводом относительно исходного запроса и аксиом глобальной онтологии. На этом шаге алгоритм производит построение альтернативных формулировок на основе позитивных аксиом терминологии исчерпывающим применением к запросу ряда правил замены предикатов, а также отсеивание конъюнкций, заведомо невыполнимых согласно негативным аксиомам терминологии. Затем к запросу применяются алгебраические методы оптимизации – проталкивание фильтров и т.п. Результатом этого этапа в терминах дескриптивной логики является объединение конъюнктивных запросов с простыми ограничениями.

На следующем этапе для каждого атома (в терминах SPARQL – для каждой RDF-тройки) каждого конъюнктивного запроса на основе конфигураций адаптеров определяется множество источников данных, которые содержат релевантные этому атому данные. Следующие правила используются для отбора релевантных источников:

- Для атомарных концептов (`?x rdf:type`) релевантные данному атому источники определяются при помощи VoID-дескрипторов наборов данных, в которых указывается, что эти источники содержат элементы класса ;
- Для атомарных ролей (`?x :predicate ?y`) релевантными будут источники, словари которых содержат предикат `:predicate` (определяется по VoID-дескрипторам, аналогично);
- Если в предыдущем случае на месте объекта или/и субъекта стоит URI ресурса, то источник будет релевантным только, если URI принадлежит этому ресурсу (оба URI принадлежат этому ресурсу).

Если для какого-либо атома нет релевантных ему источников данных, то весь конъюнктивный запрос отбрасывается. В результате этого мы получаем объединение конъюнктивных запросов с атомами из различных наборов данных.

Традиционно для систем интеграции данных по принципу посредников следующим шагом является составление оптимального физического плана исполнения запроса и собственно его исполнение. При исполнении конъюнктивного подзапроса с атомами из разных источников происходит объединение результатов подзапросов к этим источникам. Однако в случае источников - независимых наборов Linked Open Data объединение результатов может произойти только по литеральным полям (что не встречается в запросах от подсистемы публикации). В нашем случае объединение результатов происходит при помощи хранимых в системе наборов связей между наборами данных источников системы. Каждый конъюнктивный запрос представляет собой шаблон графа, в вершинах которого расположены литеральные значения, либо URI ресурсов, либо переменные, а ребра помечены предикатами в терминах различных источников данных. Если два смежных ребра помечены предикатами из разных источников, то необходимо обратиться к наборам связей идентичности (`owl:sameAs`) между этими двумя источниками и отобрать множество пар ресурсов, удовлетворяющих связи между двумя наборами данных. Прделав такую операцию для всех связей в запросе, получим множество ресурсов, которые удовлетворяют части шаблона графа, определяющей связи между наборами данных.

Затем части запроса, относящиеся к отдельным источникам данных, независимо исполняются адаптерами, причем участвовавшие в связи вершины заменяются на соответствующие URI из наборов связей, а результаты этих независимых подзапросов объединяются. В терминах SPARQL такое возможно при помощи конструкции BINDINGS, введенной в версии языка 1.1. Возможно параллельное исполнение частей запроса различными источниками, либо же последовательное исполнение, в случае наличия фильтров в подзапросах к источникам. Полученные от источников результаты соединяются, затем ответы на все альтернативные конъюнктивные запросов объединяются в конечный результат.

9. Связывание данных

Согласно рекомендациям проекта Linked 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. Важным типом связей являются связи идентичности, которые указывают на то, что два URI синонимичны и используются для идентификации одного и того же объекта реального мира или абстрактного понятия. Связи идентичности могут быть использовать такие предикаты как owl:sameAs, rdfs:seeAlso или специальные SKOS [22] термины. Хотя использование предиката owl:sameAs в пространстве LOD зачастую противоречит семантике OWL, его использование в пространстве LOD рекомендуется W3C Technical Architecture Group. При этом связь следует рассматривать не как факт в терминах дескриптивной логики, а как утверждение некоторого поставщика данных.

Подсистема интеграции данных занимается установлением и поддержкой связей идентичности, найденные связи помещаются в подсистему хранения. Сохраняется информация о степени идентичности ресурсов, источнике связи и связываемых наборах данных. На Рис.4. представлена схема данных хранилища связей:

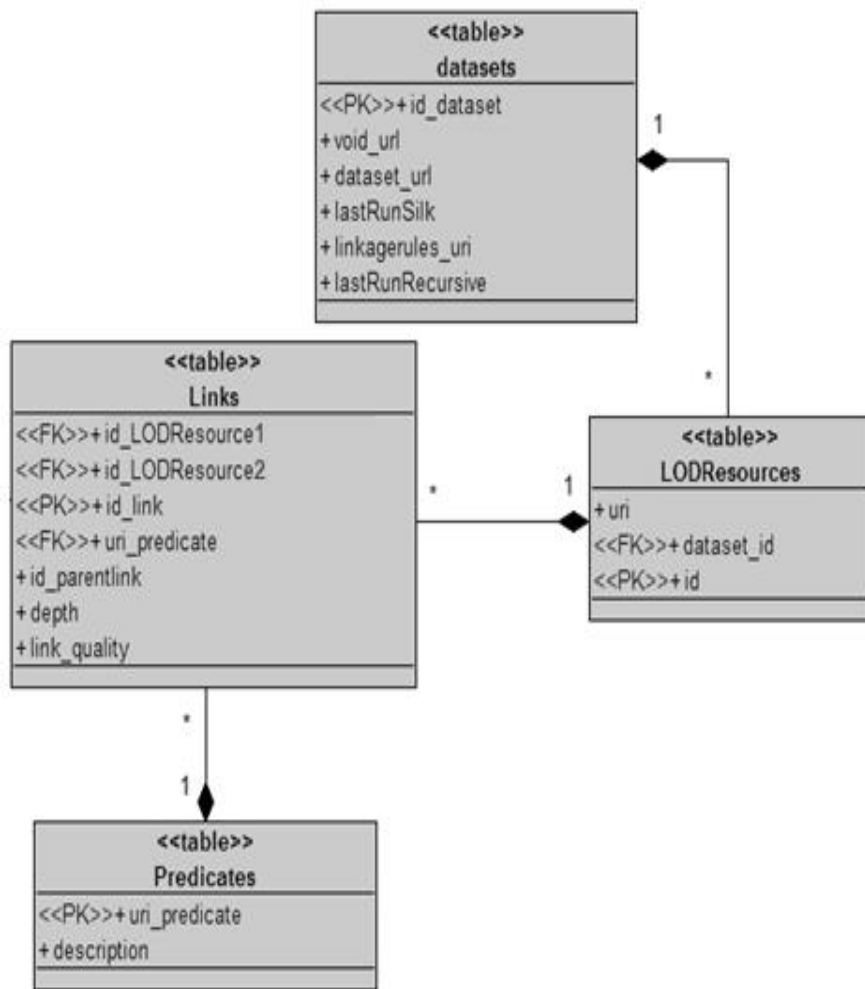


Рис. 4 Схема хранилища связей

Подсистема связывания данных работает следующим образом. На первом этапе определяются два набора данных, между элементами которых необходимо установить связи. Информация о выборе пар наборов данных для связывания помещается в подсистему хранения и затем публикуется в виде наборов связей VoID. При этом набор связей изначально пуст. При подключении нового источника данных к системе автоматически создаются наборы связей между этим новым набором данных и всеми существующими наборами данных других источников. Набор связей между одним из источников данных системы и внешним набором данных LOD может быть создан в одном из следующих случаев:

- Пользователь системы вручную указывает пару наборов данных для связывания
- Следуя по связям во внешних наборах данных, а также по уже установленным связям, подсистема связывания обнаружила два набора связанных через цепочку других наборов данных

После того, как были выбраны два набора данных для связывания, подсистема должна кластеризовать наборы данных по классам и определить пары кластеров для связывания. Это делается с целью уменьшения числа попарных сравнений элементов наборов данных. В случае двух внутренних наборов данных системы оба описываются онтологией системы, поэтому кластеры образуют экземпляры

одного класса онтологии в этих наборах данных. В случае связывания со внешним набором данных подсистема связывания может определить пары классов при помощи различных техник сопоставления онтологий [11][6], а также при помощи найденных либо указанных вручную правил отображения онтологий [2].

Наконец, на третьем этапе подсистема связываний должна попарно сопоставить элементы кластеров из двух наборов данных и определить, какие из них являются идентичными. Для этого используются правила языка SILK LSL. В случае внутренних источников данных системы правила разрабатываются совместно с онтологией и определяют, какие экземпляры одного и того же класса считаются идентичными. В случае внешнего набора данных правила могут быть также специфицированы вручную, либо же выведены из существующих правил связывания и правил отображения.

10. Заключение

В статье предлагается проект системы интеграции пространственных данных. Прикладной областью применения системы является работа с данными об ООПТ, однако подход может быть легко применен для любых пространственных данных. Система позволяет публиковать данные в традиционных форматах (GML, SHP) в пространстве Linked Open Data, а также устанавливать связи между ними и открытыми данными из внешних источников. Необходимость генерации ссылок позволяет применять в системе новый комбинированный подход к федеративным SPARQL запросам. Система может применяться для интеграции данных из многочисленных автономных источников, между которыми могут быть выявлены достаточно стабильные связи. Направления дальнейших работ включают:

- Исследование вопроса автоматической генерации онтологии, схемы реляционной базы данных и правил отображения между ними по прикладной GML схеме;
- Добавление поддержки других распространенных форматов обмена геоданными (KML, GeoRSS и т.п.);
- Дальнейшую оптимизацию метода ответов на поисковые запросы при помощи наборов связей;
- Исследование вопросов автоматического обнаружения релевантных наборов данных, оптимального выбора пар ресурсов для связывания;
- Добавление поддержки работы с пространственными типами данных в SPARQL-запросы

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., Schultz A. The R2R Framework: Publishing and Discovering Mappings // In First International Workshop on Consuming Linked Data, 2010.

- [3] Calvanese D., De Giacomo G., Lembo D., Lenzerini M., Poggi A., Rodriguez-Muro M., Rosati R., Ruzzi M., Savo D. F. The MASTRO system for ontology-based data access // Semantic Web Journal, volume 2, number 1, pages 43-53, 2011.
- [4] D2R Server: Accessing databases with SPARQL and as Linked Data [Электронный ресурс] // The D2RQ Platform [сайт] URL: <http://www4.wiwiwiss.fu-berlin.de/bizer/d2r-server/> (дата обращения: 20.06.2012).
- [5] Dublin Core Metadata Element Set [Электронный ресурс] // DCMI Home: Dublin Core® Metadata Initiative (DCMI) [сайт] URL: <http://dublincore.org/documents/dces/> (дата обращения: 20.06.2012).
- [6] Euzenat J., Ferrara A. First results of the ontology alignment evaluation initiative 2011. // In Proc. of 6th Ontology Matching Workshop, 2011.
- [7] GeoNames [Электронный ресурс] URL: <http://www.geonames.org/> (дата обращения: 20.06.2012).
- [8] GeoSPARQL - A geographic query language for RDF data [Электронный ресурс] // World Wide Web Consortium (W3C) [сайт] URL: <http://www.w3.org/2011/02/GeoSPARQL.pdf> (дата обращения: 20.06.2012).
- [9] GeoSpecies Knowledge Base [Электронный ресурс] URL: <http://about.geospecies.org/> (дата обращения: 20.06.2012).
- [10] Gorlitz O., Staab S. SPLENDID: SPARQL Endpoint Federation Exploiting VOID Descriptions // Proceedings of the 2nd International Workshop on Consuming Linked Data, Bonn, Germany, 2011.
- [11] Gracia J., Bernad J., Mena E. Ontology matching with CIDER: Evaluation report for the OAEI 2008 // In Proc. of 3rd Ontology Matching Workshop, 2008.
- [12] Heath T., and Bizer C. Linked Data: Evolving the Web into a Global Data Space (1st edition) // Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool, 2011, pp. 1-136. URL: <http://linkeddatabook.com/editions/1.0/> (дата обращения: 20.06.2012).
- [13] INSPIRE Data Specification on Geographical Names [Электронный ресурс] // INSPIRE Infrastructure for Spatial Information in European Community [сайт] URL: http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_GN_v3.0.pdf (дата обращения: 20.06.2012).
- [14] INSPIRE Data Specification on Protected sites [Электронный ресурс] // INSPIRE Infrastructure for Spatial Information in European Community [сайт] URL: http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_PS_v3.0.pdf (дата обращения: 20.06.2012).
- [15] INSPIRE Guidelines for the encoding of spatial data [Электронный ресурс] // INSPIRE Infrastructure for Spatial Information in European Community [сайт] URL: http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_GN_v3.0.pdf

nspire.jrc.ec.europa.eu/reports/ImplementingRules/DataSpecifications/D2.7_v3.0.pdf (дата обращения: 20.06.2012).

[16] NeoGeo Geometry Ontology [Электронный ресурс] // GeoVocab.org [сайт] URL: <http://geovocab.org/geometry.html> (дата обращения: 20.06.2012). [17] Ngonga Ngomo A.C., Auer S. LINES - a time-efficient approach for large-scale link discovery on the web of data // Proceedings of the Twenty-Second international joint conference on Artificial Intelligence, Vol. 3, pp. 2312-2317.

[18] Nikolov A., d'Aquin M. Identifying Relevant Sources for Data Linking using a Semantic Web Index // Workshop: 4th Workshop on Linked Data on the Web (LDOW 2011) at 20th International World Wide Web Conference (WWW 2011), Hyderabad, India.

[19] OGR: OGR Simple Feature Library - GDAL [Электронный ресурс] // GDAL - Geospatial Data Abstraction Library [сайт] URL: <http://www.gdal.org/ogr/> (дата обращения: 20.06.2012).

[20] OWL Web Ontology Language [Электронный ресурс] // World Wide Web Consortium (W3C) [сайт] URL: <http://www.w3.org/TR/owl-features/> (дата обращения: 20.06.2012).

[21] RDF Primer [Электронный ресурс] // World Wide Web Consortium (W3C) [сайт] URL: <http://www.w3.org/TR/rdf-primer/> (дата обращения: 20.06.2012).

[22] SKOS Simple Knowledge Organization System Primer [Электронный ресурс] // World Wide Web Consortium (W3C) [сайт] URL: <http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/> (дата обращения: 20.06.2012).

[23] SPARQL Query Language for RDF [Электронный ресурс] // World Wide Web Consortium (W3C) [сайт] URL: <http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/> (дата обращения: 20.06.2012).

[24] State of the LOD Cloud [Электронный ресурс] // Freie Universitat Berlin [сайт] URL: <http://www4.wiwi.fu-berlin.de/locloud/state/> (дата обращения: 20.06.2012).

[25] Tschirner S., Scherp A., Staab S. Semantic Access to INSPIRE - How to Publish and Query Advanced GML Data // Proceedings of the Terra Cognita Workshop on Foundations, Technologies and Applications of the Geospatial Web, 2011, pp. 75-87.

[26] Van der Plaats M. Dataset linkage recommendation on the Web of Data (Master thesis), 2011.

[27] Virtuoso Universal Server [Электронный ресурс] // OpenLink Software [сайт] URL: <http://virtuoso.openlinksw.com/> (дата обращения: 20.06.2012).

[28] 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.

[29] W3C Semantic Web Interest Group: Basic Geo [Электронный ресурс] // World Wide Web Consortium (W3C) [сайт] URL: <http://www.w3.org/2003/01/geo/> (дата обращения: 20.06.2012).

[30] Кузнецов К.А. Исследование и реализация распределенного поиска на основе онтологий. М: Московский Государственный Университет им. М.В.Ломоносова, 2009, 54 с.

[31] Федеральный закон. № 33-ФЗ «Об особо охраняемых природных территориях» от 14 марта 1995 г. [Электронный ресурс] // Информационно-правовое издание Legis [сайт] URL: <http://www.legis.ru/misc/doc/312/>.

Об авторах

Кузнецов Константин Александрович – аспирант МГУ им. М.В. Ломоносова, e-mail: K.Kuznetcov@gmail.com

Серебряков Владимир Алексеевич – Вычислительный Центр РАН e-mail: serebr@ccas.ru

Теймуразов Кирилл Борисович – Вычислительный Центр РАН e-mail: kbt@ccas.ru
