On Geographical Binding of the Content of Text Documents

Oleg L. Zhizhimov, Yulia V. Leonova
Institute of Computational Technologies SB RAS, Novosibirsk

Abstract. Extracting geographical names from arbitrary text documents is important in the tasks of processing large arrays of documents and linking their content to a specific geographic region. In the simplest form, the model for extracting geographical names from the text looks like a sequence of actions with the text, while at each stage its task is solved. Among these tasks, there are undoubtedly: text parsing, analyzing text elements, processing synonyms and abbreviations, bringing the text elements to normal form from possible word forms and grammar rules, comparing text elements with the elements of dictionaries of geographical names, adding special tags to the text for unambiguous identification geographical names. The proposed work describes a technology that implements the above tasks on the basis of a freely distributed PostgreSQL DBMS. In this case, the standard configuration is used, all the server part settings are performed within the framework of the documented procedures. GeoNames Gazetteer database, Open Street Map (OSM) databases, ОКАТО and КЛАДР classifications are used as an authoritative database of geographical names.

Keywords: geographical names, full-text search, model of extraction of names, text processing, PostgreSQL, geographical search.

О ГЕОГРАФИЧЕСКОЙ ПРИВЯЗКЕ КОНТЕНТА ТЕКСТОВЫХ ДОКУМЕНТОВ

Жижимов О.Л.[(1)], Леонова Ю.В.[(1)]

[(1)] Институт вычислительных технологий СО РАН, г. Новосибирск

Извлечение географических названий из произвольных текстовых документов имеет важное значение в задачах обработки больших массивов документов и привязки их контента к определенному географическому региону. В самом простом виде модель извлечения географических названий из текста выглядит как последовательность действий с текстом, при этом на каждом этапе решается своя задача. Среди этих задач, несомненно, присутствуют: парсинг текста, анализатор элементов текста, обработка синонимов и сокращений, приведение элементов текста к нормальной форме с возможных словоформ и правил грамматики, сравнение элементов текста с элементами словарей географических названий, добавление в текст специальных меток для однозначной идентификации географических названий. В предлагаемой работе описана технология, реализующая перечисленные выше задачи на базе свободно распространяемой СУБД PostgreSQL. При этом используется стандартная конфигурация, все настройки серверной части выполнены в рамках штатных документированных процедур. В качестве авторитетной базы данных географических названий применены база данных GeoNames Gazetteer, базы данных Open Street Map (OSM), классификаторы ОКАТО и КЛАДР.

Ключевые слова: географические названия, полнотекстовый поиск, модель извлечения названий, обработка текста, PostgreSQL, географический поиск.

Введение. Цель настоящей работы - создание модели извлечения географических названий из произвольного текста с его индексирование по географическим атрибутам, например, по географическим координатам, с возможностью дальнейшей организации геометрического поиска.

Следует отметить, что существующие программные комплексы для организации доступа к текстовым информационным ресурсам не обладают необходимой функциональностью по хранению и обработке географических данных. Наделение же их требуемой функциональностью осложняется отсутствием единых стандартов на поиск и представление данных, связанных с географическим аспектом, которые сопрягались бы с существующими геоинформационными системами (ГИС), т. е. с системами, для которых географический аспект информации является основным [1]. Отсюда вытекает актуальность и перспективность создания технологии, обеспечивающей обработку географической информации в «негеографических» информационных системах общего назначения [2].

Модель и алгоритмы. Если очень коротко описать предлагаемую модель фиксирования географического контента в текстовом массиве данных для последующей индексации, то она будет выглядеть следующим образом.

1. Первое, что необходимо сделать при обработке произвольного текста - раскрыть все сокращения. В тексте заменяются слова-сокращения на их несокращенные значения. Эта процедура существенна для дальнейшего анализа, т.к. в текстах географические названия как правило сопровождаются сокращенными обозначениями типа географического объекта: г. - город, оз. - озеро, обл. - область и т.п. При этом необходима не только простая механическая подстановка значений в соответствии со словарем сокращений, но и анализ сопутствующего контента. В частности, сокращение «г.» может восприниматься те только как «год», но и как «город», в зависимости от окру-

жающих слов. Формализованные правила, в соответствии с которыми происходит раскрытие сокращений, образуют специальный словарь шаблонов сокращений.

2. Полученный в результате вышеописанной процедуры текст разбивается на отдельные слова (токенизация) с фиксацией порядкового номера каждого слова в исходном тексте. При этом также происходит удаление стоп-слов, определенных в специальном словаре, и приведение остальных слов к нормальной форме в соответствии с морфологическим словарем, который может сводить множество разных лингвистических форм слова к одной лексеме.
3. Следующий желательный, но не обязательный шаг, - раскрытие перечислений. Дело в том, что в различных текстах часто встречаются различные перечисления географических названий с групповым указанием типа объекта. Например, текст «... исследования были проведены в Новосибирской, Кемеровской и Омской областях» для однозначной фиксации географических объектов требует его преобразования к виду «... исследования были проведены в Новосибирской области, Кемеровской области и Омской области».
4. После выполнения вышеперечисленных процедур можно выполнить фиксацию географических объектов - приписать специальные метки соответствующим комбинациям слов или заменить соответствующую комбинацию слов на специальную метку. Первый вариант необходим в случае дальнейшей индексации текста как для геометрического, так и для полнотекстового поиска, а второй - для индексации географических объектов только для поиска геометрического.

Географическая фиксация текстов осуществляется путем поиска в тексте географических названий из географического словаря, в котором кроме написания этих названий указываются принадлежность к географической реальности: городу, поселку, горе, болоту, реке, морю и т.д. Текст документа сравнивается с шаблонами, содержащих общие географические термины – слова, которые определяют характер географического объекта, его род и вид (например, город, гора, озеро). Как говорилось выше, географическое название может представлять собой последовательность из нескольких слов, содержать тире, часть слов или все слова могут быть написаны с заглавной буквы, также в тексте могут идти подряд несколько названий. Задача извлечения географического названия состоит в определении границы этого названия (последовательности слов, в него входящей) в окрестности географического термина. Географические названия в тексте имеют разнородную внутреннюю структуру. Частью автоматической обработки является выявление этой структуры. Можно считать, что географические названия образуются по определенным формулам – словообразовательным моделям. Для выявления географических в тексте необходимо учитывать их правила написания в русском языке [10], используемые для задания конструкций лексико-семантических шаблонов, которые определяют входящие в конструкцию слова с учетом их морфологических характеристик.

Выявленные названия должны проверяться на правильность написания в соответствии с правилами грамматического согласования. Для определения начальной формы географического названия, чтобы избежать омонимии - совпадения в некоторых падежных формах, необходимо определить род, число и падеж анализируемых слов. Общие географические термины выполняют роль родовых слов, с которыми согласуются географические названия.

При определении принадлежности лексем к словарю рассматривается термин и его окрестность из двух слов, что позволяет находить названия, состоящие из одного или двух слов, употребляющихся в сочетании с общим географическим термином, например, город Великие Луки.

Удобным средством извлечения географических названий из теста являются лингвистические шаблоны. Лингвистический шаблон содержит формальное описание (образец) языковой конструкции, которую необходимо найти в тексте, чтобы извлечь название. Шаблон может быть записан с использованием регулярных выражений и учитывать особенности слов: регистр букв, последовательности букв, например, ([А-Я] [а-я]+) край

Специальная метка может представляет собой уникальный идентификатор географического объекта в базе данных географических названий. Формально вся процедура сводится к замене нормализованных лексем на специальные метки с идентификатором объекта или на метки с лексемами. Соответствие лексем и меток содержится в специальном географическом словаре.

5. Наконец, последний шаг - разрешение проблемы многозначности географических названий. Например, вполне определенной форме «Советский район» может быть поставлено в соответствие более 40 географических объектов (на основе данных [5]). Однако, среди всех возможных нужно выбрать тот, который максимально соответствует окружающему контексту. Здесь возможны несколько вариантов разрешения конфликта.
   - На основе иерархических связей - решение об идентификации объекта среди конкурирующих принимается на основе анализа иерархических связей соседних по тексту полностью идентифицированных объектов. Иерархические связи (административное подчинение, географическое расположение и пр.) как правило присутствуют в базах данных географических названий. Более того, идентификаторы объектов некоторых баз данных хранят эту иерархию в значении идентификационного кода, например, справочник ОКАТО [3]. В частности, для города Карасук код ОКАТО 50217501 содержит информацию о Карасукском районе (ОКАТО 50217000) и Новосибирской области (ОКАТО 50000000).
   - На основе геометрических параметров - решение об идентификации объекта среди конкурирующих принимается на основе минимизации расстояния до соседних по тексту полностью идентифицированных объектов. Расстояние вычисляется на основе координат объектов, присутствующих в базе данных географических названий. При этом возможны различные варианты критерия принятия решения.

Алгоритм фиксации географических объектов в произвольном тексте изображен на Рис. 1.

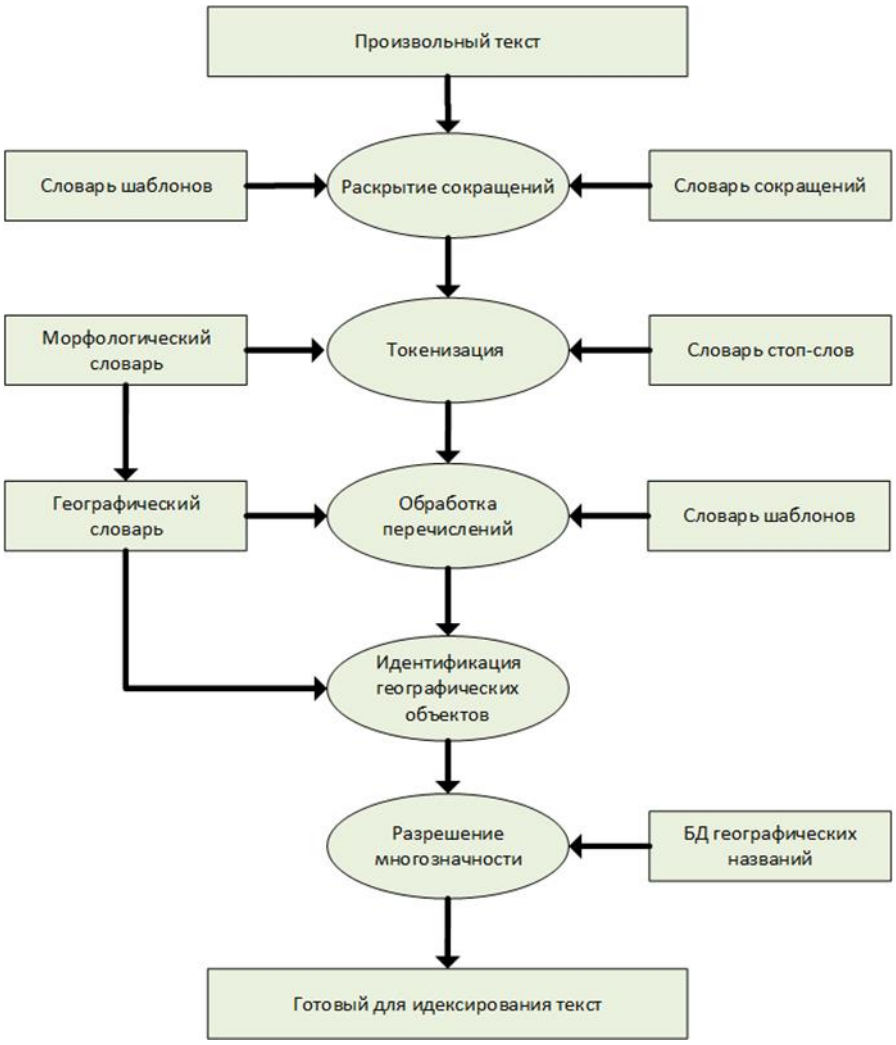

Рисунок 1. Алгоритм фиксации географических объектов в произвольном тексте

Справочники. Как следует из вышеописанной технологии, качество ее работы зависит от качества и полноты справочников, содержащих информацию о географических названиях. Сегодня в открытом доступе существует достаточно информации для создания собственной базы данных географических названий. Источниками информации могут быть:

- ОКАТО - общероссийский классификатор объектов административно-территориального деления [3].
- КЛАДР - класификатор адресов Российской Федерации [4].
- GeoNames - база данных, содержащая свыше 10 млн. географических названий и информацию о более 7,5 млн их уникальных характеристик. Среди характеристик: названия мест на разных языках, широта, долгота, высота над уровнем моря. Все эти характеристики разбиты по категориям, так что каждая характеристика географического объекта относится к одному из девяти классов. А каждая из этих категорий, в свою очередь, делится на подкатегории, общее количество которых более 600. Кроме наименований на различных языках, хранятся географические координаты, высота над уровнем моря, численность населения, административное деление

и почтовые индексы. К сожалению, база данных содержит дубли, ошибки в наименованиях и другие неточности.

- База данных OSM (Open Street Map) [6] - открытая база данных географических объектов, включающая их геометрические и географические характеристики
- Getty тезаурус географических названий (TGN) [7] - содержит географические названия с точечными координатами, в том числе и ретроспективные. Недостаток - российские названия даны в транскрипции.
- Государственный каталог географических названий РОСРЕЕСТР [8] - содержит полный реестр официальных географических названий по регионам с точечными координатами.

Перечисленные информационные ресурсы содержат исходные данные, на основе которых формируется собственная база данных географических объектов, описанная ниже.

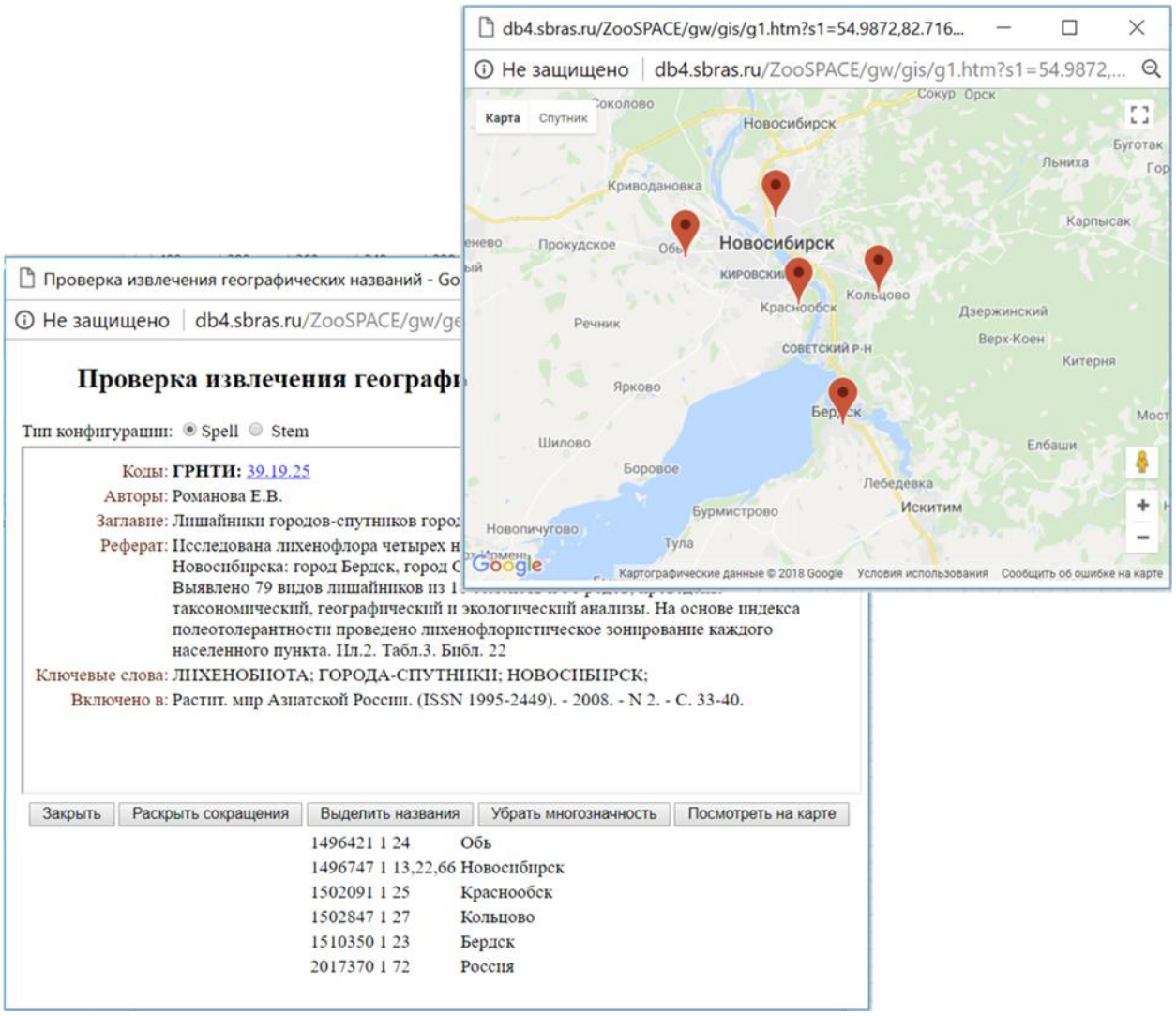

Рисунок 2. Интерфейсы приложения для тестирования алгоритмов

Прототип стенда. Для отработки технологии извлечения географических названий из текстов, проведения тестирования алгоритмов и сбора информации об ошибках был создан программный стенд, в котором реализованы описанные выше алгоритмы. В качестве системной основы реализации алгоритмов был выбран вариант на базе СУБД PostgreSQL, реализующей полный цикл обработки текстовой информации с возможностью расширения базовых функциональных возможностей как за счет дополнительных словарей и конфигура-

ций, так и написанием дополнительных модулей на различных языках программирования [9].

Созданный прототип стенда включает в себя:

- Базу данных СУБД PostgreSQL 9.4 со специальной конфигурацией для полнотекстового поиска, ориентированного на географические объекты.
- Набор серверных WEB приложений (PHP скрипты), которые выполняются на стороне WEB сервера. Эти приложения обеспечивают связь с сервером баз данных PostgreSQL и клиентскими приложениями. Отдельным серверным приложением также является модуль для ZooSPACE, позволяющий анализировать текстовые данные, извлекаемые из различных библиографических баз данных.
- Набор клиентских WEB приложений (Java скрипты), которые выполняются на стороне WEB клиента. Эти приложения реализуют функции графических интерфейсов пользователя для управления работой стенда и визуализации найденных географических объектов на картах.

Для обеспечения работы стенда созданы

- Словари
  - Словарь сокращений с шаблонами на основе регулярных выражений - при помощи этого словаря раскрываются сокращения во входном тексте (шаг 1).
  - Словарь стоп-слов русского языка (russian.stop). Этот словарь входит в поставку PostgreSQL и в нашей конфигурации не менялся (шаг 2).
  - Морфологический словарь русского языка (ispell) с добавлением географических названий и орфографических правил для этих названий (ru_geo1.dict).

    Фрагмент ru_geo1.dict

    ```
    . . .
    абажур/K
     . . .
    Кольцово/M
    Мошковский/A
    Новосибирск-Южный/AEZ
    . . .
    ```

  - Географический словарь для замены лексем на комбинацию «метка + лексема». Этот словарь (geor1.ths) соответствует шаблону тезауруса (в терминах PostgreSQL тезаурус - это словарь замен: левая часть от символа «:» заменяется на правую часть, наличие символа «*» в первой позиции правой части предписывает не контролировать правую часть морфологическим словарем) и состоит из записей вида:

    Фрагмент файла geor1.ths

    ```
    . . .
    Бердск: */gn/1510350 Бердск
    город Бердск: */gn/1510350 город Бердск
    Советский район: */gn/490026,/gn/1491227 Советский район
    . . .
    ```

- Конфигурация FTS (в терминах PostgreSQL), определяющую список словарей и порядок обработки текста (rugeo1):

  Команды создания конфигурации FTS rugeo1

  ```
  CREATE TEXT SEARCH DICTIONARY rugeo_ispell (TEMPLATE = ispell,
      dictfile = 'ru_geo1', afffile = 'ru', stopwords = 'russian');
  ```

```
CREATE TEXT SEARCH DICTIONARY tz_geo_1 (TEMPLATE = thesaurus,
    dictfile = 'geor1', dictionary = 'rugeo_ispell');
CREATE TEXT SEARCH CONFIGURATION rugeo1 (PARSER = "default");
ALTER TEXT SEARCH CONFIGURATION rugeo1 ADD MAPPING
    FOR hword WITH tz_geo_1,rugeo_ispell,russian_stem;
ALTER TEXT SEARCH CONFIGURATION rugeo1 ADD MAPPING
    FOR hword_part WITH tz_geo_1,rugeo_ispell,russian_stem;
```

Работу алгоритма фиксации географических названий можно проиллюстрировать на примере обработки фрагмента текста «В окрестностях города Новосибирска находятся: город Бердск, город Обь, поселок Краснообск и Наукоград Кольцово». В результате выполнения запроса

```
SELECT plainto_tsquery('rugeo1', 'В окрестностях города Новосибирска находятся:   город Бердск, город Обь, поселок Краснообск и Наукоград Кольцово');
```

получим ответ - размеченный текст

```
'окрестность /gn/1496747 город новосибирск находиться
/gn/1510350 город бердск /gn/1496421 город обь /gn/1502091 поселок
краснообск /gn/1502847 наукоград кольцово'
```

Другой запрос

```
SELECT to_tsvector('rugeo1',
'В окрестностях города Новосибирска находятся: город Бердск, город Обь,
поселок Краснообск и Наукоград Кольцово');
```

вернет список лексем с указанием их позиции в тексте

```
"'/gn/1496421':10 '/gn/1496747':3 '/gn/1502091':13
'/gn/1502847':17 '/gn/1510350':7 'бердск':9 'город':4,8,11
'кольцово':19 'краснообск':15 'наукоград':18 'находиться':6
'новосибирск':5 'обь':12 'окрестность':2 'поселок':14"
```

Вид интерфейсов приложения для тестирования алгоритмов на стенде представлен на Рисунке 2.

Заключение. В результате выполненных работ был создан прототип стенда для тестирования моделей и алгоритмов извлечения географических названий из неструктурированного текста для построения индексов как для текстового, так и для геометрического поиска. Предварительное тестирование показало, что предложенная технология обеспечивает высокую степень достоверности результатов при условии, что все справочники содержат информацию об определяемых географических объектах. Эффективность технологии зависит от полноты справочников. В настоящее время созданные справочники содержат информацию по географическим объектам Новосибирской области. В дальнейшем планируется расширить номенклатуру поддерживаемых регионов.

Работа выполнена при частичной финансовой поддержке Интеграционного Проекта СО РАН (АААА18-118022190008-8), проекта по фундаментальным научным исследованиям (АААА-А17-117120670141-7), проекта РФФИ № 18-07-01457-а, а также при частичной поддержке грантового финансирования научных и (или) научно-технических исследований на 2018-2020 гг. МОН РК (№ АР 05133546).

ЛИТЕРАТУРА

1. Жижимов О. Л., Мазов Н. А. Проблемы географической привязки цифровых объектов в электронных библиотеках. Тр. XII Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL’2010). Казань, с. 207–214. (2010).
2. Барахнин В.Б., Жижимов О.Л., Куперштох А.А., Скачков Д.М., Федотов А.М. Алгоритм извлечения из текстовых документов географических названий, отражающих содержание. Вестник Новосибирского государственного университета. Серия: Информационные технологии, Т.10, № 1, С.109-120. (2012).
3. Общероссийский классификатор объектов административно-территориального деления (ОК 019-95), http://protect.gost.ru/document.aspx?control=20&id= 134377.
4. Класификатор адресов Российской Федерации (КЛАДР), http://kladr-rf.ru.
5. The GeoNames geographical database, http://www.geonames.org//
6. Open Street Map, http://wiki.openstreetmap.org.
7. Getty Thesaurus of Geographic Names (TGN), http://www.getty.edu/research/ tools/vocabularies/tgn/index.html.
8. Государственный каталог географических названий, РосРеестр, https://rosreestr.ru/site/activity/ geodeziya-i-kartografiya/naimenovaniya-geograficheskikh-obektov/ gosudarstvennyy-katalog-geograficheskikh-nazvaniy/.
9. Бартунов О., Сигаев Ф. Введение в полнотекстовый поиск в PostgreSQL, http: //citforum.ru/database/postgres/fts/bib.shtml.
10. Основные правила написания географических названий http://www.wikiznanie.ru/ru-wz/index.php/Основные_правила_написания_географических_названий