Географический поиск в информационных системах с использованием ретроспективного тезауруса

© Д. М. Скачков © О. Л. Жижимов
Институт вычислительных технологий Сибирского отделения РАН,
г. Новосибирск
danil.skachkov@gmail.com zhizhim@sbras.ru

Аннотация

В статье рассматривается задача поиска в информационных системах по географическому аспекту и, в частности, поиск с использованием тезауруса ретроспективного геокодирования. Обсуждаются вопросы, связанные с особенностями географической привязки цифровых объектов. Приводится вариант реализации географического поиска в информационной системе, а также результаты экспериментальной интеграции.

Выполнено при частичной поддержке СО РАН (IV.31.1.1, ИП-2012-17, ПИП-2012-73), РФФИ (10-07-00302-а, 12-07-00472-а), Президиума РАН (Проекты 2012-14.3, 2012-15.2), ФЦП шифр номер 2012-1.4-07-514-0022-004

1 Географический аспект в цифровых объектах

До середины 1960 годов карты являлись всего лишь способом хранения символьной информации о географических объектах. 1960-е годы были ознаменованы появлением географических информационных систем или ГИС. ГИС это информационная система, обеспечивающая сбор, хранение, обработку и визуализацию пространственных данных и связанной с ними информации. Уже тогда было заявлено, что приоритетной задачей картографии является не создание визуальных продуктов, а процессы сбора, преобразования и обработки информации. И основаны эти процессы будут на компьютерных системах [1]. Сегодня, за счет того что технологии шагнули далеко вперед, географические данные стали широко доступны. И за счет интернет сервисов, таких как Google Maps™[10], стало возможным интегрировать функциональность ГИС в системы, которые для этого не были предназначены изначально. Это так называемые «негеографические» информационные системы, к которым относятся, например, электронные каталоги, базы данных научно-технической информации, архивы с информацией о цифровых и нецифровых объектах. Но тот факт, что эти системы не были предназначены для работы с географической информацией, еще не говорит, что эта информация там не содержится. Любая статья была где-то написана и опубликована, любой экспонат музея был где-то найден, тексты научных трудов зачастую содержат названия географических объектов. И это только несколько примеров того, что «негеографические» системы на самом деле содержат географическую информацию.

Можно выделить следующие типы систем, информация в которых потенциально имеет географическую компоненту: системы хранения информации о цифровых и физических объектах (системы хранения метаданных) и системы хранения собственно цифровых объектов. При этом первые отличаются от вторых только формализованной структурой данных и формализованной семантикой наполнения. Ниже мы не будем делать различия между этими системами, в основном рассматривая подсистему метаданных, которые с необходимостью присутствуют в обоих типах систем после каталогизации цифровых объектов. Для определенности, рассматриваемые информационные системы будем относить к классу электронных библиотек.

Традиционные правила каталогизации физических и цифровых объектов предписывают создание метаданных, ориентированных на структуру и семантику стандартизованных схем данных. Для библиографической информации такими схемами являются (MARC21, RUSMARC, МЕКОФ и др.). Географический аспект объектов содержится в полях, которые связаны с некоторым географическим местом, временной аспект – в полях, которые связаны с некоторым событием.

Географический и временной аспекты в описании объекта, как правило, связаны, т.к. любое описанное событие характеризуется временем и местом.

В качестве примеров каталогизируемой информации, содержащих данные о событиях, в библиографических массивах данных можно указать:

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

• контент, т.е. информационное содержание объекта (ключевые слова, аннотации, текст и пр.)

• события создания объекта (выполнение работы, съемка, написание, перевод и пр.)

• события публикации (издание, переиздание и пр.)

• события хранения (помещение в репозиторий, музей, библиотеку и пр.)

• события проведения мероприятия (конференции, выставки, реставрация и пр.)

• и др.

2 Географическая привязка в информационной системе

2.1 Ретроспективное геокодирование

К сожалению, прямое использование географического аспекта каталогизированных в соответствии с действующими правилами каталогизации событий для географического поиска неэффективно [8]. Дело в том, что географическая информация хранится в текстовых полях и пригодна только для простейшего поиска по географическому названию. Такой поиск существенно не отличается от обычного текстового поиска. Но поиск по географическому аспекту информации имеет свои отличительные особенности.

Рассмотрим следующую задачу: необходимо найти все научные статьи, которые были опубликованы на территории Новосибирской области. При этом мы не можем просто произвести поиск по словосочетанию «Новосибирская область», т.к. с одной стороны, в соответствии с правилами каталогизации в метаданных содержится только название города, а с другой, - в данном географическом регионе находится множество других объектов: города Новосибирск, Бердск, Барабинск, Карасук и множество других населенных пунктов. Таким образом, чтобы найти все релевантные статьи мы должны составить список из всех населенных пунктов Новосибирской области, и производить поиск по каждому из них. Более того, некоторые населенные пункты, существовавшие в прошлом, в настоящее время не существуют или были переименованы. Таким образом, к нашему списку населенных пунктов мы должны добавить еще и населенные пункты, существовавшие в прошлом, а также устаревшие названия населенных пунктов. Все становится еще сложнее, если необходимо найти материалы, в которых упоминаются объекты новосибирской области, т.е. не только населенные пункты, но и реки, озера, улицы, железнодорожные станции и подобные им объекты. Составить такой список вручную будет практически невозможно.

Одним из решений данной проблемы является географическая привязка объектов информационной системы. Под географической привязкой мы будем понимать логическую связь цифрового объекта с некоторой геометрической областью на земной поверхности. При наличии такой привязки в информационных объектах электронной библиотеки, задача поиска объектов, релевантных заданному региону, сводится к простейшей задаче проверки перекрытия геометрических областей, выполняемой математическими методами, а не методами, основанными на лексическом анализе. Такой подход обеспечивает, во-первых, большую релевантность результатов, по сравнению с текстовым поиском по названиям географических объектов, и, во-вторых, такая функциональность уже встроена во многие хранилища данных, чего нельзя сказать об алгоритмах лексического анализа. Большая релевантность результатов, в данном случае, следует из однозначности географических координат. Если мы будем производить текстовый поиск по запросу «Алексеевка» (имея в виду деревню Алексеевка Московской области), то получим большое количество результатов не относящихся к нашему запросу, поскольку населенных пунктов с названием «Алексеевка» в России великое множество, и названием «Алексеевка» нельзя однозначно идентифицировать географический объект. В то же время, произведя поиск по географическим координатам 55°47′12.52″ с. ш. 38°18′33.02″ в. д. мы получим только результаты, относящиеся к искомому географическому объекту, в данном случае деревне Алексеевка в Ногинском районе Московской области.

Способы географической привязки объектов были подробно рассмотрены в [8-9]. Здесь и ниже мы будем рассматривать способ привязки посредством тезауруса [12]. Такая привязка осуществляется с помощью добавления к записям системы идентификатора или идентификаторов объектов из соответствующего тезауруса. При этом осуществляется привязка некоторой информации, содержащей место и время, т.е. ассоциированной с некоторым событием. Поэтому в рамках задачи географического поиска наиболее целесообразно использовать тезаурус ретроспективного геокодирования, описанный в [11]. Тезаурус ретроспективного геокодирования отличается от других тезаурусов географических наименований наличием информации об изменениях состояния географических объектов с течением времени. Таким образом, учитывая, что в информационных системах зачастую хранятся данные относящиеся к прошедшим моментам времени, причем достаточно отдаленным, видим, что только из тезауруса ретроспективного геокодирования мы можем получить наиболее достоверные данные о состоянии географических объектов во время определенных событий.

2.2 Индексация существующих данных

Естественно, наиболее интересна реализация событийного географического поиска для уже

существующих информационных массивов и систем. При использовании тезауруса эта процедура достаточно проста: необходимо добавить в структуру записей базы метаданных информационной системы поля для хранения географических идентификаторов записей и проиндексировать все записи идентификаторами терминов, входящих в тезаурус географических наименований. При индексации следует учесть, что данные в электронных библиотеках могут содержать не только единичные упоминания географических объектов, но и множественные. Поэтому поля для хранения идентификаторов объектов из тезауруса должны позволять хранить как один элемент, так и множество.

Индексация данных информационной системы производится с помощью алгоритма, описанного в [4]. Приведем основные этапы данного алгоритма индексации. Первым этапом решения поставленной задачи является извлечение из текста документа всех географических названий, входящих в тезаурус. Фактически, мы имеем дело с задачей координатного индексирования текста терминами, входящими в заданный словарь, при этом термины могут состоять не только из одного, но и из нескольких (как правило, двух) слов, например, Новосибирская область, Белое море, Северная Двина и т.п. В [13] описан алгоритм автоматического поиска и подсчета ключевых слов из заданного словаря, представляющих собой словосочетания сложной структуры, учитывающий морфологию русского языка. В основу алгоритма [5] положено использование двух индексов, содержащих триады

«*номер текста*» – «*позиция в тексте*» – «*номер слова из лексического словаря*»

и

«*номер термина*» – «*позиция слова в термине*» – «*номер слова из лексического словаря*».

При этом если первый индекс существует практически во всех информационно-поисковых системах, то введение второго индекса, позволяющее резко повысить эффективность алгоритма, имеет оригинальный характер. Индекс терминов наряду с их списком размещается в хранилище данных программной библиотеки, реализующей алгоритм, и пополняется по мере изменения этого списка.

Кратко опишем указанный алгоритм.

I. Алгоритм построения индекса терминов состоит из следующих этапов:

1. Разбиение термина на отдельные слова.

2. Создание предварительного индекса, содержащего триады «номер термина» – «позиция слова в термине» – «слово в символьном представлении».

3. Добавление встретившихся неизвестных слов в лексический словарь библиотеки, где им присваиваются идентификационные номера.

4. Переработка индекса в формат «номер термина» – «позиция в тексте» – «номер слова из лексического словаря».

5. Сбор статистики о длинах терминов для реализации поиска и идентификации составных терминов (т.е. терминов, состоящих более чем из одного слова).

6. Сбор статистики о количестве вхождений отдельных слов в термины для оптимизации поиска путем исключения из рассмотрения терминов, заведомо отсутствующих в тексте.

II. Алгоритм построения индекса текстов аналогичен, но в нем отсутствует этап 3.

III. Заключительная стадия работы программной библиотеки – подсчет количества вхождений терминов в текст (тексты). Ее этапы:

1. Подсчет возможных комбинаций «текст» – «термин», основанный на статистике вхождения отдельных слов (см. этап 6 алгоритма индексации терминов).

2. Нахождение всех потенциально возможных мест вхождения каждого термина в текст (тексты) на основе наличия хотя бы одного общего слова из лексического словаря. Позиция каждого потенциально возможного вхождения фиксируется.

3. Рассмотрение каждого из возможных мест вхождений с точки зрения соответствия термину в целом. Актуальность вхождения определяется наличием рядом с соответствующей позицией других слов, входящих в термин. Существуют конфигурируемые варианты требований определения актуальности вхождения (точный или неточный порядок слов, минимальное количество слов, входящих в термин, возможность «прерывания» термина посторонними словами и т. п.).

4. Исключение учета вхождений, поглощаемых более длинными вхождениями.

5. Сбор статистики вхождений для каждой пары «текст» – «термин».

Отметим, что при решении задачи извлечения географических названий этапы 3 и 4 актуальны довольно редко, но все-таки их нельзя полностью исключить: например, практически равноупотребительны термины Новосибирский район и Новосибирский сельский район, обозначающие один и тот же географический объект.

Напрямую использовать тезаурус географических наименования в данном алгоритме не получится, так как при анализе текстов необходимо учитывать морфологию русского языка. Словарь должен быть пополнен словоформами географических наименований. Автоматическая генерация словоформ может быть осуществлена посредством использования библиотеки морфологического анализа phpMorphy [7]. Однако ранее проведенные с ней эксперименты по генерации словоформ математических терминов, входящих в тезаурус предметной области «Математика» [6], показали

высокую, но отнюдь не стопроцентную правильность генерации словоформ. Поэтому в тех случаях, кто алгоритмы библиотеки дают неправильный результат, следует прибегать к непосредственной генерации словоформ экспертом.

К сожалению, существует еще одна проблема, усложняющая задачу извлечения из текста документа географических названий. Дело в том, что географические названия могут быть омонимичны другим словам, являющимися именами как нарицательными: Орёл, Белая и т.п., так и собственными: Киров, Кострома и т.п. Кроме того, нередко одно и то же название носят сразу несколько различных географических объектов. Возникает необходимость отсеять из полученного набора слов омонимы географических названий, таковыми не являющиеся, а также установить, к какому конкретно географическому объекту относится найденное в документе «многозначное» название.

Для выявления в тексте омонимов географических названий, и для конкретизации значения «многозначных» названий, необходимо заранее в процессе работы с тезаурусом составить список географических названий, имеющих такие омонимы, и список «многозначных» названий. Если «многозначные» названия в тезаурусе выявляются достаточно просто, путем его непосредственного анализа, то выявление омонимов «общего плана» - задача более сложная. Наиболее общим приемом выявления нарицательных омонимов является учет регистра первой буквы слова. Этот прием может оказаться неэффективным, если омонимичное слово является первым словом в предложении, а также если заголовок документа набран прописными буквами. В случае неоднократного вхождения такого слова в текст почти наверняка удастся выявить его смысл путем анализа регистра первой буквы всех его вхождений. Если же омонимичное слово встречается только раз и притом в качестве первого слова в предложении, то относить его к географическим названиям вряд ли целесообразно хотя бы потому, что географические названия зачастую употребляются с предлогом указания места или направления (т.е. не выступают в качестве первого слова предложения), а в случае возможной омонимии – и с указанием типа географического объекта (город Орёл, река Белая и т.п.) [4].

Результатом приведенного алгоритма будет список из идентификаторов записей нашего тезауруса. После записи полученных идентификаторов объектов в соответствующие поля, база метаданных информационной системы будет готова для географического поиска.

2.3 Алгоритм поиска

Рассмотрим простейший способ реализации поиска в информационной системе, записи которой проиндексированы географическими идентификаторами (Рисунок 1):

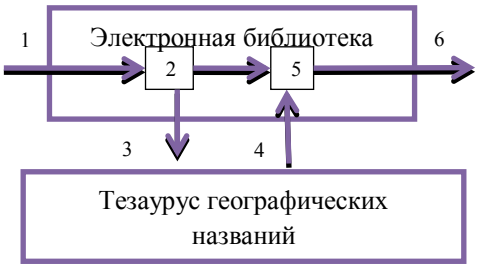

Рис. 1. *Сценарий географического поиска в электронной библиотеке*

1) передача поискового запроса в информационную систему;

2) выделение части поискового запроса, относящейся к географическому поиску;

3) передача географического запроса в тезаурус;

4) получение из тезауруса списка идеентификаторов географических объектов, релевантных данному поисковому запросу;

5) формирование запроса информационной системы на основе исходного запроса и полученных идентификаторов географических объектов и его выполнение;

6) возврат результата.

Пример 1 (запросы RPN в нотации PQF, определения CIP из [3]): Найти все записи в базе данных, которые соответствуют ресурсам, опубликованным в Новосибирской области с 12 октября 2001 года по 10 января 2007 года (текстовое представление)

@and

@attr 1=59 @attr 2=3 @attr 4=108 {Новосибирская область}

@attr 1=31 @attr 2=16 @attr cip 4=210 {2001-10-12,2007-01-10}

Пример 2 (запросы RPN в нотации PQF): Найти все записи в базе данных, которые соответствуют ресурсам, опубликованным в Новосибирской области с 12 октября 2001 года по 10 января 2007 года (геометрическое представление)

@and

@attr 1=59 @attr cip 2=7 @attr cip 4=202

{((53.3590,75.2152),(57.2273,85.1248))}

@attr 1=31 @attr cip 2=16 @attr cip 4=210 {2001-10-12, 2007-01-10}

Рассмотрим подробнее этапы выполнения запросов:

1. Передача запроса в информационную систему производится посредством интерфейса пользователя. Если для поиска текстовых данных интерфейс ввода параметров существует в виде текстового поля и успешно используется достаточно давно, то для составления геометрического запроса необходим интерфейс, позволяющий выбирать области на географической карте. Такой интерфейс

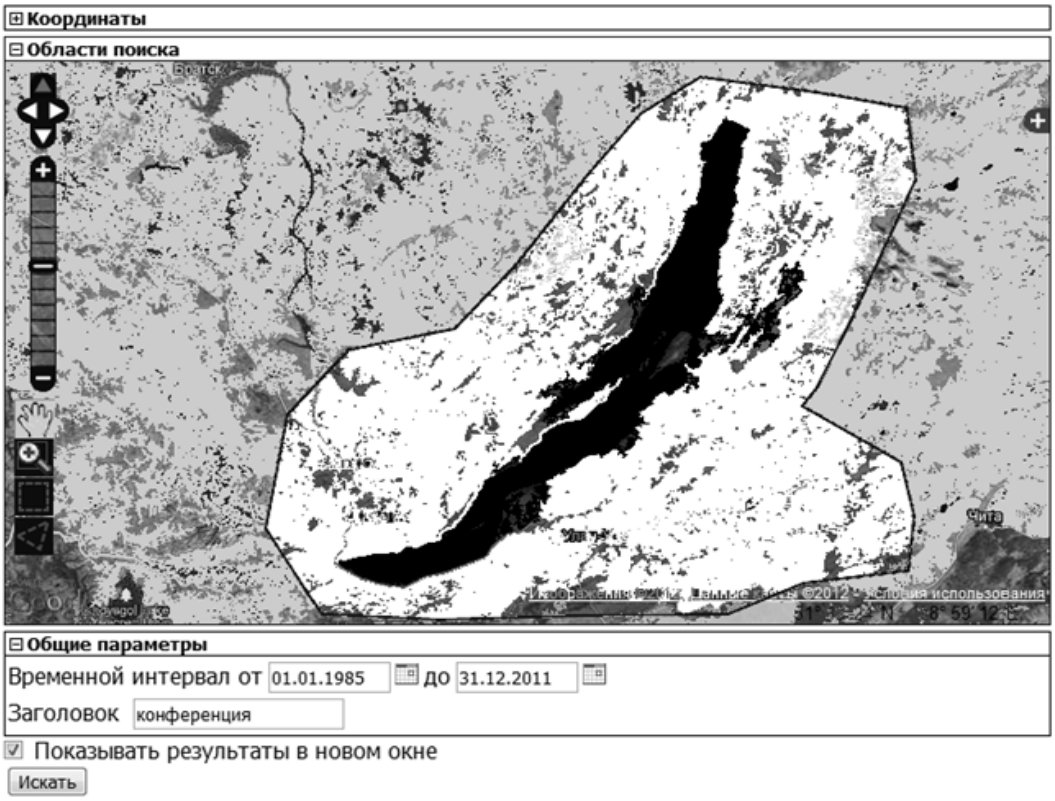

Рис. 2 *Интерфейс создания поискового запроса*

предоставляет, например, сервис Google Maps [2]. Используя данный сервис можно реализовать необходимый пользовательский интерфейс (Рисунок 2).

Следует заметить, что в пользовательском интерфейсе формирования запроса должна быть возможность указания временного интервала для искомых событий.

2. Поскольку поисковый запрос может содержать не только географическую часть, на втором этапе географическую и временную части следует выделить для обработки во внешней подсистеме, реализующей сервисы информационной системы тезауруса.

3. Тезаурус принимает на вход поисковый запрос со следующими параметрами:

- тип и название региона (пример 1) или тип и координаты географического региона (пример 2);
- временной период события (необязательный).

4. Тезаурус в ответ на запрос возвращает список идентификаторов объектов, находящихся в заданном географическом регионе. Если в параметрах поиска был указан временной интервал события, то возвращается список объектов, находящихся в заданном регионе в указанный период времени. Например, идентификаторы населенных пунктов, Новосибирск, Барабинск, Черепаново …

5. С использованием полученного списка идентификаторов географических объектов и параметров из оригинального запроса, формируется запрос к библиографической базе данных. В результате его выполнения мы получаем список релевантных объектов. В приведенных примерах запрос к библиографической базе данных будет выглядеть следующим образом (для обоих примеров):

@and
 @and @and . . .
@attr 1=59 @attr 2=3 @attr 4=108 {Новосибирск}
@attr 1=59 @attr 2=3 @attr 4=108 {Черепаново}
@attr 1=59 @attr 2=3 @attr 4=108 {Барабинск}
. . .
@and
 @attr 1=31 @attr 2=4 @attr 4=5 {20011012}
 @attr 1=31 @attr 2=2 @attr 4=5 {20070110}

Заметим, что при этом запрос сформирован в терминах набора Bib-1, который обычен для поиска библиографической информации.

6. Система возвращает список найденных объектов и отображает их в интерфейсе пользователя.

Следует уточнить способы доступа к тезаурусу. В нашем случае тезаурус представляет собой базу данных, доступ к которой осуществляется по протоколам

- Z39.50,
- HTTP/XML/SOAP/SRW,
- HTTP/SRU,

Таблица 1. Некоторые точки доступа профиля RGeoThes

Точка доступа	**Набор**	**Тип**	**Значение**
Тип определения геометрического объекта (точка, полигон)	cip-1	4	201, 202
Координаты геометрического объекта	cip-1	1	2059, 2060
	cip-1	2	7,8,9,10
Тип определения события (временной интервал)	cip-1	4	210
Временной интервал события	cip-1	1	2062
	cip-1	2	14,15,16,17,18

согласно профилю доступа RGeoThes, определенному в [12]. Профиль определяет ряд точек доступа к данным, находящимся в тезаурусе. В рамках данной задачи интересны следующие точки доступа (Таблица 1).

Схемы и форматы извлечения записей из базы данных тезауруса соответствуют стандартным спецификациям упомянутых протоколов, например формат XML, схема ZThes.

Таким образом, рассмотренная выше технология использования тезауруса позволяет существенно расширить поисковые возможности «негеографических» информационных систем в область геометрического географического поиска с использованием графических пользовательских интерфейсов, основанных на картографических сервисах.

Таблица 2. Результаты поиска с применением географического тезауруса

Заголовок	**Год публикации**
Международная конференция "Почва как связующее звено функционирования природных и антропогенно-преобразованных экосистем", Иркутск, 2-6 сентября 2006	2007
Международная конференция "Ультрамафит-мафитовые комплексы складчатых областей докембрия" на Байкале п. Энхалук, 6-9 сент., 2006	2007
Международная конференция по охране озера Байкал	2004
В Иркутске состоялась международная конференция "Управление земельными ресурсами с особым акцентом на защиту окружающей среды в районе озера Байкал"	2006
Международная конференция по экологии Сибири, пос. Листвянка, 24-27 августа 1993 г.	1994
В Иркутске состоялась международная конференция "Управление земельными ресурсами с особым акцентом на защиту окружающей среды в районе озера Байкал"	2006
Молодежная научная конференция по органической химии "Байкальские чтения 2000", Иркутск, 18-25 июля, 2000	2000
Третья международная конференция "Энергетическая кооперация в Северо-Восточной Азии: предпосылки, условия, направления", Иркутск 9-13 сент., 2002 г	2003
Евроазиатская авиатранспортная научно-практическая конференция "Аэропорты Сибири и Дальнего Востока. Потенциал роста", Иркутск, 30 июня, 2005, проводимая в рамках 4 Байкальского экономического форума, Иркутск, 2005	2005
12 Байкальская международная конференция "Методы оптимизации и их приложения", Иркутск, 24 июня - 1 июля, 2001	2001
14 Байкальская международная школа-семинар "Методы оптимизации и их приложения" и 3 Всероссийская научная конференция "Равновесные модели экономики и энергетики", Северобайкальск, 2-8 июля 2008	2008
13 Байкальская Всероссийская конференция "Информационные и математические технологии в науке и управлении (ИМТ 2008)", Иркутск-Байкал, 7-17 июля 2008	2008
12 Байкальская Всероссийская конференция "Информационные и математические технологии в науке, управлении, (ИМТ'2009)", Иркутск, июнь 2009	2009

3 Экспериментальная интеграция

В качестве эксперимента, была произведена интеграция географических метаданных в базу данных публикаций по исследованиям Байкальской природной зоны. Интерфейс формирования поискового запроса представлен на рисунке 2. Задав примерную область байкальской природной территории, и указав ключевое слово в заголовке «конференция» и временной интервал поиска с 1985 г. по 2011 г. мы получаем список всех записей, относящихся к данному региону и содержащих в заголовке слово «конференция» (Таблица 2).

Теперь произведем поиск без использования географических метаданных. В данном случае, мы должны искать по текстовому запросу: «Байкальская природная зона» и «конференция». Но т.к. словосочетание «Байкальская природная зона» в наименованиях не содержится вообще, то мы будем искать по словам «Байкал» и «конференция». В итоге получаем следующие результаты поиска (Таблица 3).

Из данного примера видно, что поиск без использования географических метаданных выдал не весь набор результатов, явно относящихся к указанному региону. Также видим, что поиск по определенным регионам на поверхности земли существенно затруднен в случае использования обычного текстового поиска – нам пришлось заменить термин «Байкальская природная зона» на более узкий термин «Байкал», чтобы найти хоть что-то. И если в данном случае такой подход помог, в силу того что большая часть конференций в целевом регионе содержит в названии слово «Байкал» в том или ином виде, то в иных случаях такой подход может не сработать.

Таблица 3. Результаты поиска без использования географических метаданных.

Заголовок	**Год публикации**
Международная конференция "Ультрамафит-мафитовые комплексы складчатых областей докембрия" на Байкале п. Энхалук, 6-9 сент., 2006	2007
Международная конференция по охране озера Байкал	2004
В Иркутске состоялась международная конференция "Управление земельными ресурсами с особым акцентом на защиту окружающей среды в районе озера Байкал"	2006
Молодежная научная конференция по органической химии "Байкальские чтения 2000", Иркутск, 18-25 июля, 2000	2000
Евроазиатская авиатранспортная научно-практическая конференция "Аэропорты Сибири и Дальнего Востока. Потенциал роста", Иркутск, 30 июня, 2005, проводимая в рамках 4 Байкальского экономического форума, Иркутск, 2005	2005
12 Байкальская международная конференция "Методы оптимизации и их приложения", Иркутск, 24 июня - 1 июля, 2001	2001
14 Байкальская международная школа-семинар "Методы оптимизации и их приложения" и 3 Всероссийская научная конференция "Равновесные модели экономики и энергетики", Северобайкальск, 2-8 июля 2008	2008
13 Байкальская Всероссийская конференция "Информационные и математические технологии в науке и управлении (ИМТ 2008)", Иркутск-Байкал, 7-17 июля 2008	2008
12 Байкальская Всероссийская конференция "Информационные и математические технологии в науке, управлении, (ИМТ'2009)", Иркутск, июнь 2009	2009

3 Заключение

В заключение заметим, что на основе описанной технологии сегодня формируется ряд информационных систем в рамках научно-исследовательских проектов Сибирского отделения РАН:

1. Интеграционный проект СО РАН 2012-17 «Создание сервисов и инфраструктуры научных пространственных данных для поддержки комплексных междисциплинарных научных исследований Байкальской природной зоны».

2. Партнерский интеграционный проект СО РАН (с ДВО РАН) 2012-73 «Современные технологии формирования информационной инфраструктуры для поддержки междисциплинарных исследований, в том числе для мониторинга природных и социальных процессов территорий Сибири и Дальнего Востока»

3. Другие проекты.

Литература

[1] Abresch J., Hanson A., Heron S., Reehling P. Integrating Geographic Information Systems into Library Services: A Guide for Academic Libraries // http://elib.sbras.ru:8080/jspui/handle/SBRAS/3362 - ISBN 978-1-59904-726-3

[2] API Карт Google - Google Maps API — Google Developers https://developers.google.com/maps/?hl=ru

[3] Catalogue Interoperability Protocol (CIP) Specification - Release B // CEOS/WGISS/ICS/CIP-B, Issue 2.4.75. - April 2005.

[4] 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 НГУ. Сер.: Информационные технологии. - 2012. - Т.10. - № 1. - С.109-120. - ISSN 1818-7900.

[5] Барахнин В.Б., Куперштох А.А. Алгоритм координатного индексирования электронных научных документов // Труды международной конференции «Вычислительные и информационные технологии в науке, технике и образовании». Казахстан, Павлодар, 20-22 сентября 2006 г. Т. I. С.228-232.

[6] Барахнин В.Б., Нехаева В.А. Технология создания тезауруса предметной области на основе предметного указателя энциклопедии // Вычислительные технологии. 2007. Т. 12. Специальный выпуск 2. С.3-9.

[7] Библиотека морфологического анализа phpMorphy. – http://phpmorphy.sourceforge.net

[8] Жижимов О.Л., Мазов Н.А. Об использовании географических координат при поиске библиографической информации // Научные и технические библиотеки. - 2009. - № 1. - С.54-60.

[9] Жижимов О.Л., Мазов Н.А. Проблемы географической привязки цифровых объектов в электронных библиотеках // XII Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL’2010 (Казань, Россия, 13.10 - 17.10.2010): Труды конференции. - Казань: Казан. ун-т, 2010. - С.207-214. - ISBN 978-5-98180-838-8.

[10] Карты Google http://maps.google.com/

[11] Скачков Д.М., Жижимов О.Л. Об интеграции географических метаданных посредством ретроспективного тезауруса // Информатика и ее применения. – 2012. – № 3. (в печати).

[12] Скачков Д.М., Жижимов О.Л. Об использовании ретроспективного геокодирования для географического поиска в электронных библиотеках // XIII Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011 (Воронеж, Россия, 19.10 - 22.10.2011): Труды конференции. - Воронеж: Издательско-полиграфический центр Воронежского государственного университета, 2011. - С.51-58. - ISBN 978-5-9273-1875-9.

[13] Шокин Ю.И., Федотов А.М., Барахнин В.Б. Проблемы поиска информации. Новосибирск: Наука, 2010.

Geographical search in information systems using retrospective thesaurus

Danil Skachkov, Oleg Zhizhimov

The problem of geographical search and search with retrospective geocoding thesaurus in information systems is discussed. Issues related to binding of geographical metadata to digital objects are also discussed. Description of geographical search algorithm implementation is included. Results of experimental implantation of geographical search into real database is presented.