Организация ретроспективного тезауруса для использования в задаче географического поиска в «негеографических» информационных системах[1]

Скачков Д. М. (danil.skachkov@gmail.com)

Институт вычислительных технологий Сибирского отделения РАН

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]. Со временем количество областей, в которых использовались ГИС, увеличивалось. Но, все же, ГИС оставались в сферах, напрямую связанных с географией. Многое изменилось с появлением географических интернет сервисов, таких как Google Maps [2]. Они дали возможность интегрировать функциональность географических информационных систем в системы, прямым образом не связанные с географией. 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. Ведь тот факт, что данные системы напрямую не связаны с географией, не означает, что географическая 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.

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 находятся и другие объекты: города Новосибирск, Бердск, Барабинск, Карасук, а также 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

[1] 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

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.

Итак, нам необходима географическая привязка объектов информационной системы. Привязка может быть осуществлена несколькими способами, подробно рассмотренными в [3]:

- с помощью количественного геометрического описания географического объекта на основе координат;
- с помощью ссылки на элемент некоторого тезауруса, включающего географические названия соответствующих объектов.

В [3] описываются причины, по которым интеграция с помощью тезауруса более предпочтительна, поэтому далее пойдет речь именно об интеграции географических метаданных посредством ссылки на тезаурус. 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. Однако, несмотря на то, что существует множество тезаурусов географических наименований [4], данная задача не так проста, как кажется. Проблема заключается в том, что географический аспект объектов, хранящихся в негеографических информационных системах, зачастую относится не к текущему моменту времени, а к моментам времени прошедшим. Однако с течением времени могут изменяться как географические названия, так и границы географических объектов. Будем называть это изменение свойств с течением времени ретроспективным аспектом информации. В то же время большинство тезаурусов содержит сведения, относящиеся только к текущему моменту времени, т.е. не учитывает ретроспективный аспект информации. Данная особенность препятствует использованию существующих тезаурусов географических наименований в подобных системах. Также существующие тезаурусы не совсем подходят для решения проблемы интеграции географических метаданных, т.к. в них координаты географического объекта чаще всего задаются в виде точки, в то время как реальные координаты объекта представляют собой далеко не точку, а, в общем случае, некоторую область. Что, конечно же, уменьшает полезность таких тезаурусов при проведении поиска. Поэтому более предпочтительным будет тезаурус, где положение объектов задано с помощью координат границ области, занимаемой объектом.

Таким образом, для использования в информационных системах общего назначения географического аспекта в его любом виде необходим справочный аппарат (тезаурус), который бы включал в себя не только географический аспект информации, но и ее временной (ретроспективный) аспект. В [3] приведены основные требования, которым должен удовлетворять тезаурус для того, чтобы его можно было использовать в задаче интеграции.

Рассмотрим вариант организации такого тезауруса. Наиболее правильно организовать доступ к тезаурусу по протоколу Z39.50, что обеспечит возможность простой интеграции с существующими информационными системами. Для организации такого доступа нам достаточно описать профиль доступа и определить правила отображения внутренней схемы данных тезауруса на данный профиль [5]. Определяемый профиль должен быть расширением профиля ZThes [6], т.к. именно профиль ZThes является стандартным для доступа к тезаурусам по протоколу Z39.50. Поисковые атрибуты для доступа по Z39.50 для обеспечения интероперабельности должны соответствовать поисковым атрибутам профиля Z-Thes из наборов zthes-1, utility, cross-domain (xd-1). Для поиска по времени и координатам должны использоваться атрибуты из набора cip-1. Аналогичное требование справедливо и для запросов CQL. Соответствие поисковых атрибутов точкам доступа представлено в Таблице 1.

Реализовав доступ по Z39.50, мы получаем также доступ по протоколам HTTP/XML/SOAP/SRW и HTTP/SRU благодаря использованию сервера ZooPARK [7]. Примеры запроса к тезаурусу, поддерживающему данный профиль, приведены в [8].

Таблица 1 Точки доступа записи RGeoThes

Точка доступа	Набор	Тип	Значение
Локальный номер	utility	1	4
Название терма	cross-domain	1	1
Квалификатор терма	zthes-1	1	1
Тип терма	zthes-1	1	2
Статус терма	zthes-1	1	7
Категория терма	zthes-1	1	6
Язык названия	utility	1	3
Дата начала действия названия	cip-1	1	2072
		2	14,15,16,17,18
Дата окончания действия названия	cip-1	1	2073
		2	14,15,16,17,18
Документ, фиксирующий название	cross-domain	1	6
Тип геометрического объекта	cip-1	4	201, 202
Координаты геометрического объекта	cip-1	1	2059, 2060
		2	7,8,9,10
Дата начала действия определения геометрии	cip-1	1	2072
		2	14,15,16,17,18
Дата окончания действия определения геометрии	cip-1	1	2073
		2	14,15,16,17,18
Документ, фиксирующий определения геометрии	cross-domain	1	6
Комментарий	cross-domain	1	4
Идентификатор связанного терма	zthes-1	1	4

Для хранения данных тезауруса будем использовать реляционную СУБД PostgreSQL, т.к. данная СУБД имеет встроенную поддержку всех необходимых типов данных (точки, линии, регионы и т.д.) и содержит большое количество стандартных функций по работе с ними.

Реляционная схема данных представлена на рисунке 1. Рассмотрим подробнее элементы реляционной схемы. Основной в данной схеме является таблица «Запись тезауруса» (далее будем ее упоминать как «главная таблица»), в которой находится список квалификаторов записей тезауруса. Строка из данной таблицы может содержать ссылку на предыдущий вариант записи и на родительскую запись. Связи между записями тезауруса содержатся в таблице «Связь между записями». Каждая связь содержит квалификаторы двух записей тезауруса, которые она, соответственно, связывает. Также связь характеризуется двумя документами (что представлено в виде внешних ключей). Один из документов - «Начальный документ» - определяет документ, в котором зафиксировано появление связи. Второй – «Конечный документ» - который может быть не указан, определяет документ, в котором зафиксировано исчезновение связи.

В таблице «Имя объекта» задаются наименования географических объектов, содержащихся в тезаурусе. Каждая из записей главной таблицы может быть связана с несколькими строками имен. В свою очередь, имя может быть связано только с одной записью из главной таблицы. Каждое из имен характеризуется собственно именем, а также типом объекта и языком. Под типом объекта понимается, например, тип населенного пункта. Каждая запись таблицы «Имя объекта» содержит идентификаторы двух документов – начального и конечного. «Начальный документ» определяет документ, в котором зафиксировано присвоение данного имени географическому объекту. «Конечный документ» определяет документ, в котором зафиксировано окончание срока действия данного имени.

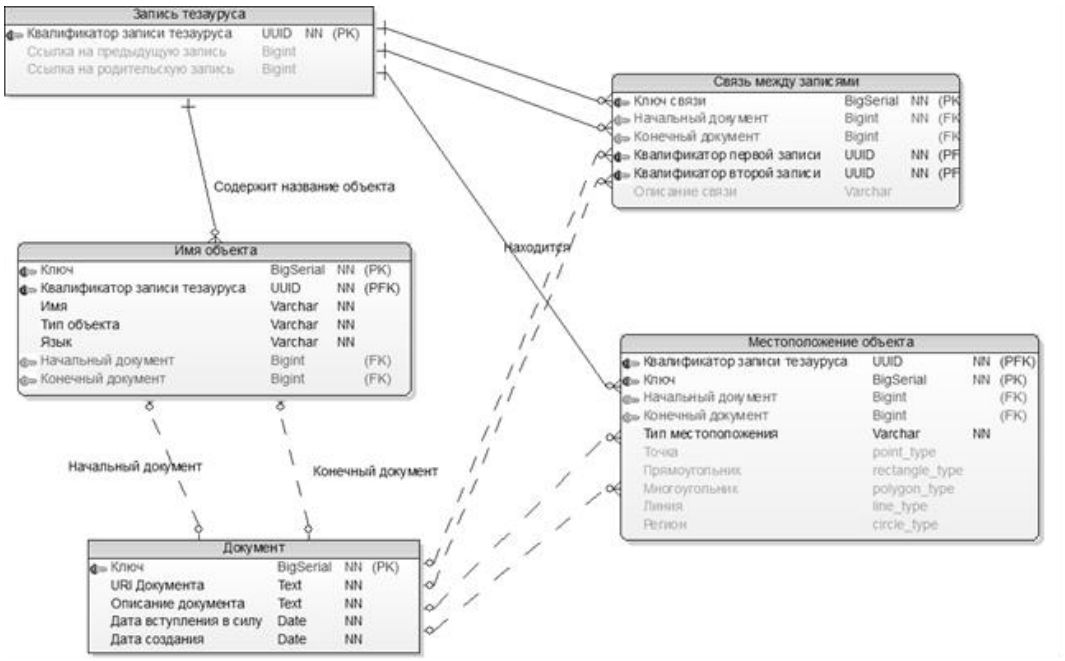

Рисунок 1 Реляционная схема данных тезауруса

Таблица «Местоположение объекта» содержит данные о координатах географических объектов тезауруса. Каждая из записей главной таблицы может быть связана с несколькими строками данной таблицы. В то же время, запись таблицы «Местоположение объекта» может быть связана только с одной строкой из главной таблицы. Каждая запись таблицы «Местоположение объекта» содержит идентификаторы двух документов – начального и конечного. «Начальный документ» определяет документ, в котором зафиксировано присвоение данного местоположения географическому объекту. «Конечный документ» определяет документ, в котором зафиксировано окончание срока действия местоположения применительно к данному объекту. Также каждая запись содержит поле «тип местоположения», содержащее идентификатор типа местоположения объекта. Таким типом может быть: точка, прямоугольник, многоугольник, линия, регион и прочие. Благодаря использованию отображения «Одна иерархия – одна таблица» для набора типов местоположений объекта становится возможным легко добавлять новые типы местоположения в уже работающую схему. Для хранения координатных данных используются поля «Точка», «Прямоугольник», «Многоугольник», «Линия», «Регион» со стандартными для PostgreSQL типами данных point, box, polygon, line, circle соответственно.

Таблица «Документ» содержит данные о документах, регистрирующих изменение характеристик объектов с течением времени. Каждый документ содержит описание, уникальный идентификатор ресурса (URI), дату создания и дату вступления в силу. Именно датой вступления в силу документов определяются временные рамки существования той или иной характеристики географического объекта.

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 записей информационной системы должна производиться по алгоритму, описанному в [9]. В основе алгоритма лежит алгоритм координатного индексирования текста терминами, входящими в заданный словарь [10]. Однако, напрямую использовать тезаурус географических наименований в данном алгоритме не получится, в силу того, что нужно учитывать морфологию русского языка. Также существует

проблема, связанная с тем, что географические названия могут быть омонимичны как друг другу, так и другим словам. Подход к решению данных проблем представлен в [9].

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

В качестве эксперимента, была произведена интеграция географических метаданных в базу данных публикаций по исследованиям Байкальской природной зоны. 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

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

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

Список литературы

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. Карты Google http://maps.google.com/
3. Скачков Д.М., Жижимов О.Л. Об интеграции географических метаданных посредством ретроспективного тезауруса // Информатика и ее применения. – 2012. – Том 6. Выпуск 3. с. 42-50.
4. 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.
5. Жижимов О.Л., Скачков Д.М. О профиле доступа к данным тезауруса для ретроспективного геокодирования и географического поиска в электронных библиотеках // XVIII Международная конференция «Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса» - Крым-2011 (Судак, Украина, 04.06 - 12.06.2011): Материалы

конференции. - М.: ГПНТБ России, 2011. - ISBN 978-5-85638-150-3. - Гос. регистр. № 0321100651. - http://www.gpntb.ru/win/inter-events/crimea2011/disk/059.pdf

6. The Zthes specifications for thesaurus representation, access and navigation. Адрес в Интернете: http://zthes.z3950.org/
7. Жижимов О. Л., Мазов Н. А. Сервер ZooPARK: вчера, сегодня, завтра // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: 14–я междунар. конф. «Крым 2007» (9–17 июня 2007 г., г. Судак): Труды конф. – М.: Изд–во ГПНТБ России, 2007. – С. 168–171.
8. Жижимов О.Л., Скачков Д.М. О географическом поиске информации в «негеографических» информационных системах: использование ретроспективного тезауруса // XIX Международная конференция «Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса» - Крым-2012 (Судак, Украина, 02.06 - 10.06.2012): Материалы конференции. - М.: ГПНТБ России, 2012. - ISBN 978-5-85638-164-0. - Гос. регистр. № 0321201404. - http://www.gpntb.ru/win/inter-events/crimea2012/disk/119.pdf
9. Барахнин В. Б., О. Л. Жижимов, А. А. Куперштох, Д. М. Скачков, А. М. Федотов. Алгоритм извлечения из текстовых документов географических названий, отражающих содержание // Вестник Новосибирского государственного университета. Серия: Информационные технологии. Том 10. Выпуск 1. - Новосибирск: Новосибирский государственный университет, 2012. - С.109-120. - ISSN 1818-7900.
10. Шокин Ю.И., Федотов А.М., Барахнин В.Б. Проблемы поиска информации. Новосибирск: Наука, 2010.