О перспективах использования тезауруса ретроспективного геокодирования в информационных системах общего назначения[1]

Скачков Д.М.[1], Жижимов О.Л.[2], Мазов Н.А.[3]
[1,2]Институт вычислительных технологий Сибирского отделения РАН,
г. Новосибирск
[3]Институт нефтегазовой геологии и геофизики им. А. А. Трофимука Сибирского отделения РАН,
г. Новосибирск
[1]danil.skachkov@gmail.com [2]zhizhim@sbras.ru [3]MazovNA@ipgg.sbras.ru

История использования географических данных в информационных системах берет начало в 1960 годах. Именно тогда становятся технически возможными и возникают так называемые географические информационные системы или ГИС. ГИС это информационная система, обеспечивающая сбор, хранение, обработку и визуализацию пространственных данных и связанной с ними информации. Уже тогда было понятно, что приоритетной задачей картографии является не создание визуальных продуктов, а процессы сбора, преобразования и обработки информации. И фундаментом для этих процессов будут компьютерные системы [1]. Сегодня, за счет того что технологии шагнули далеко вперед, географические данные стали доступны для широкого круга задач. И за счет интернет сервисов, таких как Google Maps™[2], 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.

Актуальность этих работ подтверждается интеграционными проектами, которые в настоящее время формируются в рамках научно-исследовательских работ 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

[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 № 2012-1.4-07-514-0022-004

междисциплинарных исследований, в том числе для мониторинга природных и социальных процессов территорий Сибири и Дальнего Востока»

Системы, информация в которых потенциально имеет географическую компоненту, мы можем разделить на два типа: 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. [3-5]. 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:

- 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.

К сожалению, прямое использование географического аспекта каталогизированных в соответствии с действующими правилами каталогизации событий для географического поиска неэффективно [6]. Дело в том, что географическая информация хранится в текстовых полях и пригодна только для простейшего текстового поиска по географическому названию. Но такой поиск может не устроить нас полностью, в силу того, что его результаты будут заведомо неточны.

Рассмотрим следующую задачу.

Необходимо найти все научные статьи, касающиеся территории Новосибирской области. Однако мы не можем просто произвести поиск по словосочетанию «Новосибирская область» в заглавии, 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. На Рис. 1 в качестве примера показан фрагмент карты

Новосибирской области с указанием на нем населенных пунктов, имеющих отношение к данной территории[7].

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. Поэтому, 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 другие географические объекты (реки, озера, улицы, железнодорожные станции и др.). Составить такой список вручную практически невозможно.

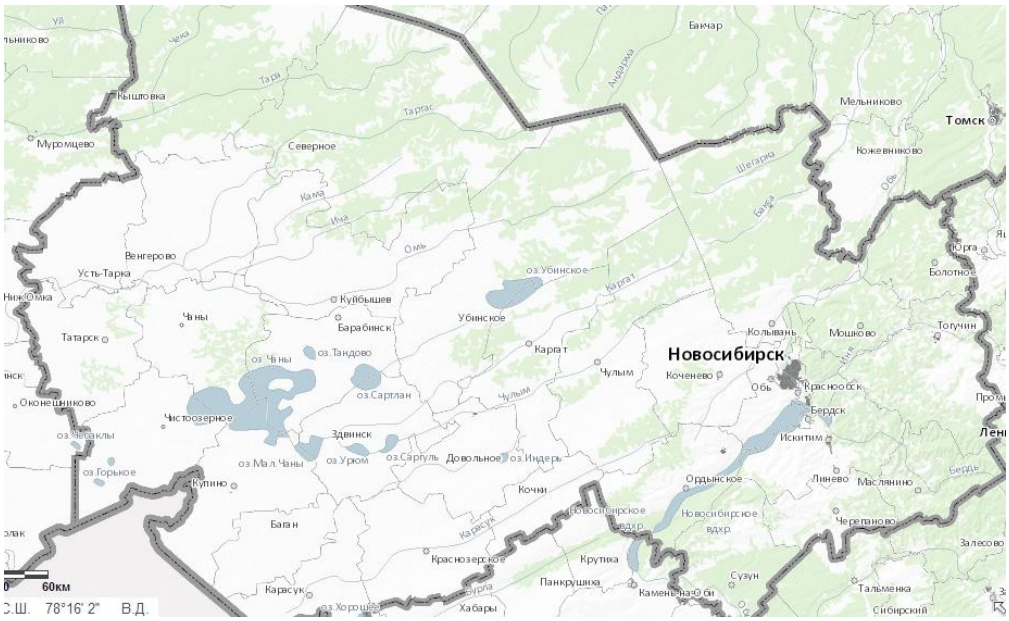

Рисунок 1. Фрагмент карты Новосибирской области

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. Такая задача выполняется математическими методами, а не методами, основанными на лексическом анализе. Данный подход обеспечивает большую релевантность результатов, по сравнению с текстовым поиском по названиям географических объектов. Кроме того, эта 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 таблице 1 представлен фрагмент справочника населенных пунктов с наименованием «Алексеевка». В то же время, произведя поиск по географическим координатам 55°47′12.52″ с. ш. 38°18′33.02″ в. д. мы получим результаты, относящиеся только к искомому географическому объекту, в данном случае деревне Алексеевка в Ногинском районе Московской области.

Таблица 1. Фрагмент справочника населенных пунктов РГБ

Название населенного пункта	**Область и район**
Алексеевка	Челябинская область, Варненский район
Алексеевка	Хабаровский край, Ванинский район
Алексеевка	Хабаровский край, Николаевский район
Алексеевка	Ульяновская область, Чердаклинский район
Алексеевка	Ульяновская область, Кузоватовский район
Алексеевка	Тамбовская область, Жердевский район
Алексеевка	Московская область, Ногинский район
Алексеевка	Саратовская область, Базарно-Карабулакский район
Алексеевка	Саратовская область, Аркадакский район
Алексеевка	Саратовская область, Красноармейский район
Алексеевка	Саратовская область, Перелюбский район

Способы географической привязки объектов подробно рассмотрены в работах [6, 8]. Здесь и ниже мы будем рассматривать способ привязки посредством тезауруса [9]. Такая привязка осуществляется с помощью добавления к записям системы идентификатора или идентификаторов объектов из соответствующего тезауруса (см. поле «qualifier» в приведенном ниже фрагменте записи). 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 [10]. Тезаурус ретроспективного геокодирования отличается от других тезаурусов географических наименований наличием информации об изменениях состояния географических объектов с течением времени (см. поле «names» в приведенном ниже фрагменте записи). 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.

Приведем фрагмент записи тезауруса (в формате JSON) для иллюстрации описанного выше:


```
{
  "qualifier":
    "7f6d7fcd-d865-4070-9182-2a8a9e464e63",

  "names": [ {
    "name": "Александровский",
    "type": "поселок",
    "language": "ru",
    "beginDocument": {
      "description": "Сход населения в октябре 1895",
```

```
      "date": "Oct 1, 1895"
    },
    "endDocument": {
      "description": "Ходатайство о переименовании поселка Александровского в поселок Новониколаевский",
      "date": "Feb 17, 1898"
    }
  },
  {
    "name": "Новониколаевский",
    "type": "поселок",
    "language": "ru",
    "beginDocument": {
      "description": "Ходатайство о переименовании поселка Александровского в поселок Новониколаевский",
      "date": "Feb 17, 1898"
    },
    "endDocument": {
      "description": "Положение Комитета министров о возведении поселка Новониколаевского в статус города",
      "date": "Jan 10, 1904"
    }
  },
  {
    "name": "Новониколаевск",
    "type": "город",
    "language": "ru",
    "beginDocument": {
      "description": "Положение Комитета министров о возведении поселка Новониколаевского в статус города",
      "date": "Jan 10, 1904"
    },
    "endDocument": {
      "description": "Постановление Президиума ЦИК СССР",
      "date": "Feb 12, 1926"
    }
  },
  {
    "name": "Новосибирск",
    "type": "город",
    "language": "ru",
    "beginDocument": {
      "description": "Постановление Президиума ЦИК СССР",
      "date": "Feb 12, 1926"
    },
  }, ],
}
```


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 [11].

Способ реализации поиска в информационной системе, записи которой проиндексированы географическими идентификаторами из тезауруса, рассмотрен в [12].Там же рассматривается способ реализации интерфейса пользователя для поиска в информационной системе по географическому региону.

Таким образом, рассмотренная 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.

В качестве эксперимента, была произведена интеграция географических метаданных в библиографическую базу данных содержащую метаописания публикаций по исследованиям Байкальской природной зоны. На рисунке 2 представлен фрагмент поискового интерфейса для формирования географического поискового запроса, построенный на основе Google Maps™.

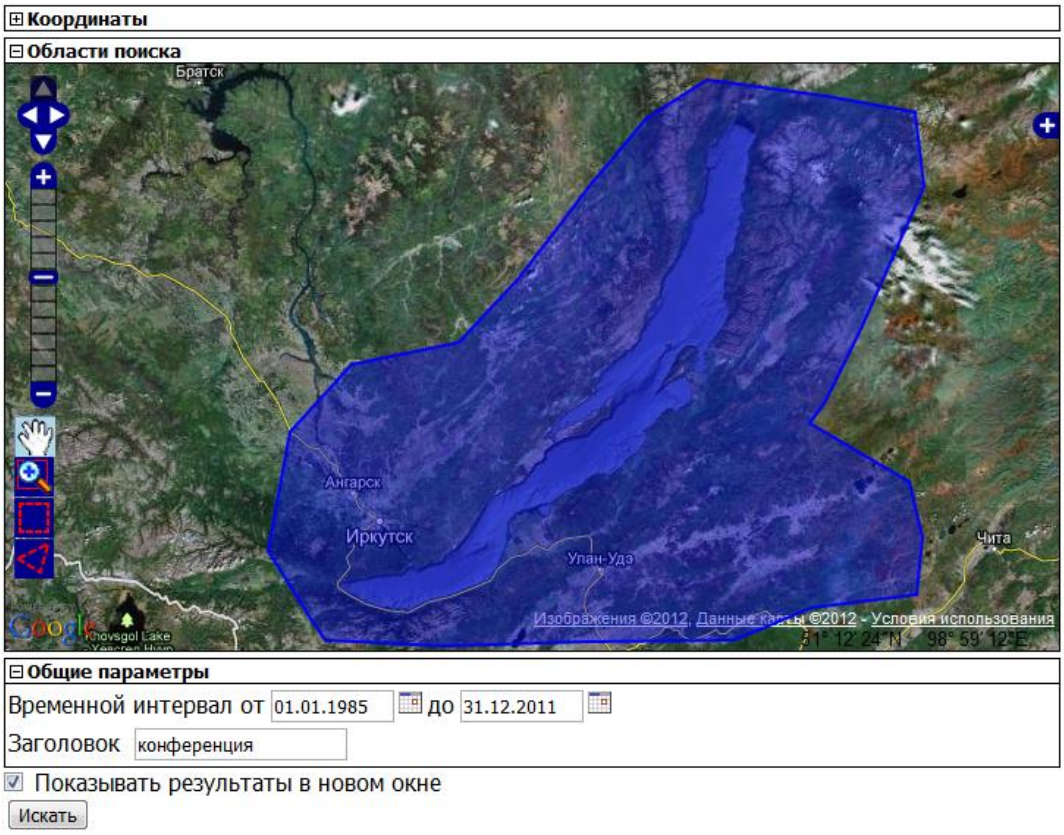

Рисунок 2. Фрагмент интерфейса построения поискового запроса.

Задав примерную область байкальской природной территории, и указав ключевое слово в заголовке «конференция» и временной интервал поиска с 1985 г. по 2011 г. производим запрос к тезаурусу.

В таблице 2 представлены результаты поиска посредством данного запроса.

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 таблице 3 представлены результаты поиска без использования географических метаданных.

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

Заголовок	Год публикации
12 Байкальская Всероссийская конференция "Информационные и математические технологии в науке, управлении, (ИМТ'2009)", Иркутск, июнь 2009	2009

Как видим, при сопоставлении результатов поиска из таблиц 2 и 3 разница составила 4 записи. В данном случае, чтобы при текстовом поиске получить те же результаты поиска, что и при поиске с использованием тезауруса, в поисковый запрос необходимо добавить географические названия «Иркутск» и «Листвянка». Из данного примера видно, что поиск без использования географических метаданных не выдал результаты, явно относящиеся 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», что не является правильным подходом.

Данный пример, к сожалению, не касается поиска ретроспективного, т.к. рассматривался достаточно небольшой период времени. Положительные стороны поиска с использованием ретроспективного геокодирования будут особенно видны при поиске по материалам большей давности.

В заключение отметим, что сама разработка технологии интеграции географических метаданных еще не дает возможности по ее использованию, поскольку одной из наиважнейших частей работы является заполнение тезауруса ретроспективного геокодирования, который используется в данной технологии. Для первоначального заполнения тезауруса предполагается использовать Справочник географических названий РГБ [13] в качестве источника данных о состоянии географических объектов. Также, для уточнения состояния географических объектов в отдаленные моменты времени будут использованы ресурсы проекта по обработке и представлению архивных карт [14 - 16]. Естественно, это не полный список источников, и по возможности, он будет расширен. Так как указанные источники не реализуют какие либо из стандартных способов доступа, извлечение информации сопряжено с определенными трудностями. Необходима разработка механизма извлечения данных для каждого конкретного источника. Однако заполнение тезауруса невозможно выполнить полностью автоматически. Для выполнения задач по проверке, исправлению и дополнению имеющихся в тезаурусе данных предполагается построение соответствующего графического интерфейса для работы с записями тезауруса.

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. Карты Google http://maps.google.com/
3. ГОСТ 7.19-2001 - Система стандартов по информации, библиотечному и издательскому делу. Формат для обмена данными. Содержание записи. // М.: ИПК Издательство стандартов.-2001.

4. RUSMARC в примерах: Учебное пособие для каталогизаторов / Национальный информационно-библиотечный центр «Либнет» - М.: Фаир-пресс. – ч.1-3. – 2003.
5. Форматы MARC21. – http://marc21.rsl.ru
6. Жижимов О.Л., Мазов Н.А. Об использовании географических координат при поиске библиографической информации // Научные и технические библиотеки. - 2009. - № 1. - С.54-60.
7. Публичная кадастровая карта. - http://maps.rosreestr.ru/PortalOnline/
8. 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.
9. 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.
10. 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.
11. 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.
12. Скачков Д. М., Жижимов О. Л. Географический поиск в информационных системах с использованием ретроспективного тезауруса // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XIV Всероссийской научной конференции RCDL'2012. Переславль-Залесский, Россия, 15-18 октября 2012 г. - г. Переславль-Залесский: изд-во "Университет города Переславля", 2012. - с. 160-167
13. Тезаурус РГБ. - http://aleph.rsl.ru/F/?func=file&file_name=find-b&local_base=tst11
14. Щекотилов В. Г. // [Электронный ресурс]. «Обработка и представление архивных карт». Режим доступа: http://boxpis.ru/ , свободный. Яз. рус. Проверено 30.10.2012.
15. Щекотилов В. Г. Создание автоматизированной интернет-коллекции по крупномасштабным топографическим межевым и военным архивным картам России XIX в. // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XIV Всероссийской научной конференции RCDL'2012. Переславль-Залесский, Россия, 15-18 октября 2012 г. - г. Переславль-Залесский: изд-во "Университет города Переславля", 2012. - с. 247-253
16. Методы обработки и совместного представления архивных и современных карт. Параллель Менде: Статьи и материалы. /Под ред. Щекотилова В.Г., - Тверь: Изд-во М.Батасовой, 2010. – 160 с.