Извлечение данных о географических объектах из существующих источников для заполнения тезауруса ретроспективного геокодирования

А.М.Тыныбекова[1], Д.Б.Мухаев[2], Д.М.Скачков[3]
[1,2]НГУ, магистрант 2 курс
[3]Институт вычислительных технологий Сибирского отделения РАН
[1]alyatynybekova@gmail.com [2]gotta_start@mail.ru
[3]danil.skachkov@gmail.com

Аннотация

В статье рассматривается метод извлечения данных о географическом объекте из системы геокодирования Google API и преобразование этих данных в соответствующий формат тезауруса ретроспективного геокодирования. Также кратко рассмотрены основные свойства географического ретроспективного тезауруса, описаны этапы извлечения данных и приведены примеры полученных записей.

Существующие на данный момент информационные системы электронных библиотек и хранилищ цифровых архивов не имеют функциональности по хранению и обработке географических данных. Однако, географическая информация содержится в записях таких систем. Причем содержится как на уровне контента, так и на уровне контекста. Добавление же функциональности, позволяющей эффективно использовать имеющуюся в записях географическую информацию, сопряжено с определенными сложностями. Реализация такой функциональности осложняется отсутствием единых стандартов на поиск и представление данных, связанных с географическим аспектом [1]. Для использования в информационных системах общего назначения географического аспекта в его любом виде необходим тезаурус, который бы включал в себя не только географический аспект информации, но и ее ретроспективный/временной (исторический) аспект [2].

Существует множество тезаурусов географических наименований, но сложность их использования заключается в том, что географический аспект объектов, хранящихся в информационных системах, зачастую относится не к текущему моменту времени, а к моментам времени прошедшим. В то время как большинство тезаурусов содержит информацию, относящуюся только к текущему моменту времени [3]. Но с течением времени могут изменяться как географические названия, так и границы географических объектов, что препятствует использованию существующих тезаурусов географических наименований в информационных системах. Изменение свойств географических объектов обычно фиксируется с помощью определенных нормативных документов.

Более того, в существующих тезаурусах 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, также уменьшает полезность таких тезаурусов при проведении поиска. Поэтому более предпочтительным будет тезаурус, где положение объектов задано с помощью координат границ области, занимаемой объектом.

Также, для задач поиска, полезными будут данные о том, как географические объекты расположены относительно друг друга. Например, если производится поиск по

некоему региону, целесообразно считать релевантными также и элементы, относящиеся к географическим объектам, лежащим в целевом регионе.

Таким образом, тезаурус должен иметь такие свойства, как:

1. наличие ретроспективных данных. Возможность извлечь данные, относящиеся к прошлому;
2. наличие связей с нормативными документами. Возможность определить, согласно какому документу было изменено название или координаты объекта;
3. описание координат географического объекта согласно его форме, т.е. географический объект может быть представлен не только в виде точки, но также и в виде замкнутого контура, линии, композиции примитивов;
4. наличие связей, отражающих относительное расположение географических объектов.

Устройство тезауруса, подходящего для использования в задаче внедрения в информационные системы общего назначения более подробно описано в [3].

Естественно, чтобы такой тезаурус можно было использовать в реальных задачах поиска, он должен быть заполнен соответствующими данными. В данном случае, тезаурус ретроспективного геокодирования должен быть заполнен информацией о географических объектах и изменении их свойств с течением времени.

Заполнение тезауруса контентом является нетривиальной задачей, так как в существующих тезаурусах (РГБ, РосРеестр, Getty) и системах геокодирования (Google API, Yandex API) данные о географических объектах хранятся в разных форматах. Для решения этой задачи необходимо создание нового механизма для заполнения тезауруса, который имел бы возможность преобразования записей из разных источников в структурированный единый формат.

Для того чтобы добавлять недостающие географические объекты в тезаурус, необходим механизм загрузки соответствующей информации из различных источников. В качестве одного из вариантов реализации такого механизма рассмотрим загрузку записей в формате XML.

Преобразование извлеченного из существующего источника XML документа с данными о географическом объекте осуществляется с помощью стилевых таблиц XSL (Рисунок 1). В этих таблицах задана структура данных, которым должен соответствовать XML документ на выходе. На вход XSLT преобразователя подаем XML документ, который был создан вручную или был выгружен из существующих тезаурусов. На выходе получаем XML документ, содержащий результат преобразования [4].

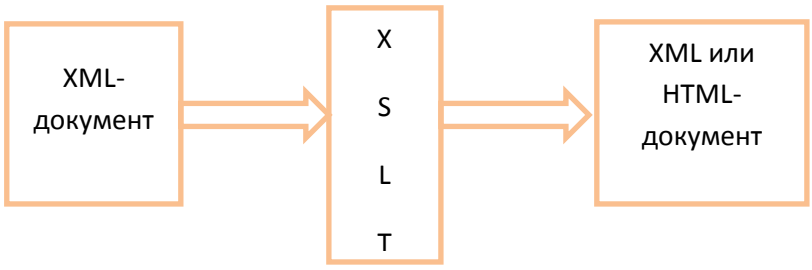

Рисунок 1 – Преобразование XML-документа

Как было сказано, XML документ может быть создан вручную или выгружен из соответствующего источника. Для задачи создания документа вручную целесообразно иметь программное обеспечение для удобного редактирования записей в XML документе. С помощью программы редактирования, можно создавать новые XML документы с описаниями географических объектов и редактировать существующие.

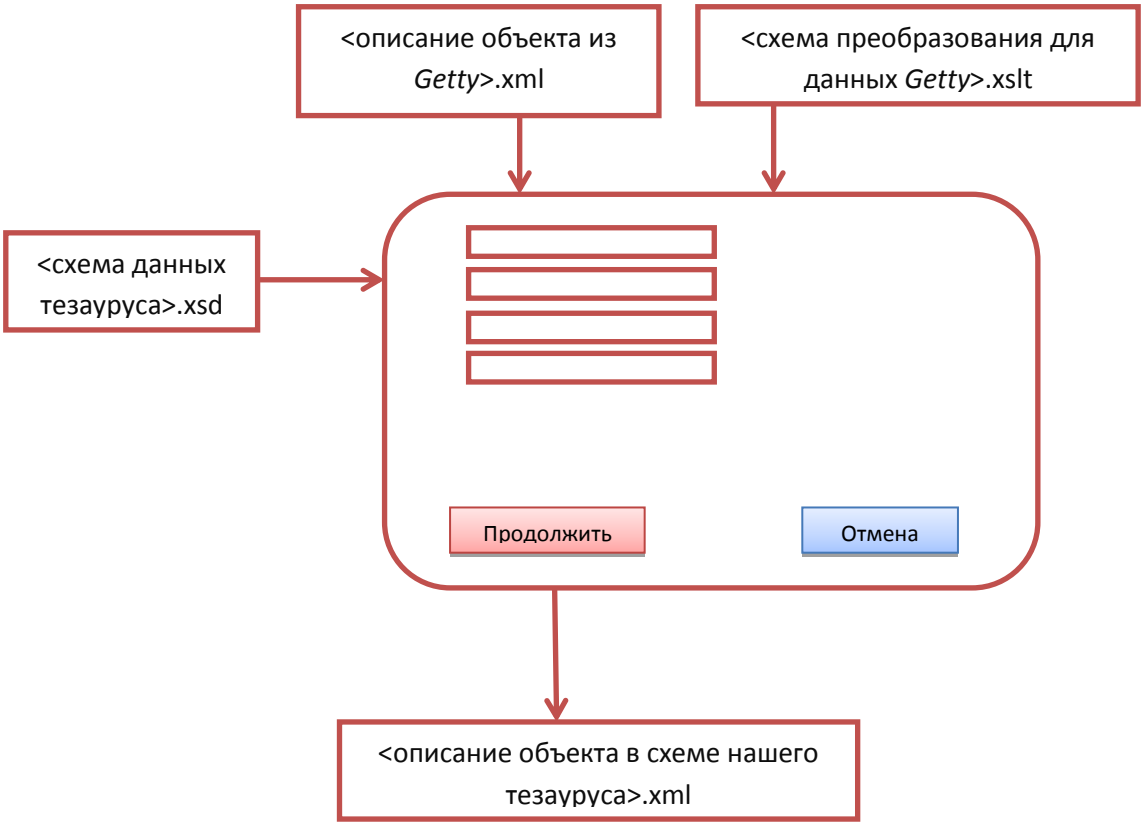

Рисунок 2 – Схема работы программы

Схема работы программы представлена на рисунке 2. Для преобразования документов с помощью механизма ему необходимо передать таблицу стилей c.xsl и схему данных b.xsd , и сам a.xml-документ, сохраненный в другом файле. Механизм произведет преобразование a.xml согласно схеме b.xsd и преобразователю x.xslt. Механизм будет читать a.xml-документ, искать в нем соответствия со схемой данных представленной в b.xsd и таблицей стилей c.xsl. В случае если механизм обнаружит соответствующие данные этим файлам в a.xml-документе и если пользователь внесет новые данные, тогда в выходной документ будут передан a'.xml-документ. А если механизм найдет какие-то несоответствия таблице стилей и схеме данных в XML-документе, то на выходной файл никакие данные не будут передаваться.

Проиллюстрируем работу описанного механизма на примере преобразования записи о географическом объекте, полученной из сервиса геокодирования Google API [5]. Запись о городе Новосибирск (1a) из Google API получаем с помощью следующего запроса: http://maps.googleapis.com/maps/api/geocode/xml?address=Новосибирск&sensor=false&languag

e=ru. К этой записи (1а) применим преобразование (1б), для разработки которого была использована программа Liquid XML Studio 2012 [6]. В результате мы получаем XML документ (1в) с описанием географического объекта. Документ в данном формате подходит для загрузки в рассматриваемый тезаурус.

1а:

<GeocodeResponse>
<status>OK</status>
<result>
<type>locality</type>
<type>political</type>
<formatted_address>Новосибирск, Новосибирская область, Россия</formatted_address>
<address_component>
<long_name>Новосибирск</long_name>
<short_name>Новосибирск</short_name>
<type>locality</type>
<type>political</type>
</address_component>
<address_component>
<long_name>Новосибирская область</long_name>
<short_name>Новосибирская область</short_name>
<type>administrative_area_level_1</type>
<type>political</type>
</address_component>
<address_component>
<long_name>Россия</long_name>
<short_name>RU</short_name>
<type>country</type>
<type>political</type>
</address_component>
<geometry>
<location>
<lat>55.0300636</lat>
<lng>82.9199423</lng>
</location>
<location_type>APPROXIMATE</location_type>
<viewport>
<southwest>
<lat>54.8035232</lat>
<lng>82.7482312</lng>
</southwest>
<northeast>
<lat>55.2003457</lat>
<lng>83.1643724</lng>
</northeast>
</viewport>
<bounds>
<southwest>
<lat>54.8035232</lat>

<lng>82.7482312</lng>
</southwest>
<northeast>
<lat>55.2003457</lat>
<lng>83.1643724</lng>
</northeast>
</bounds>
</geometry>
</result>
</GeocodeResponse>

1б:

```
<?xml version="1.0" encoding="utf-8"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
   xmlns:msxsl="urn:schemas-microsoft-com:xslt" exclude-result-prefixes="msxsl">
   <xsl:output method="xml" indent="yes"/>

   <xsl:template match="/">
     <html>
       <body>
         <h2>RGeoThes</h2>
         <table border="1">
           <tr bgcolor="#9acd32">
             <th>name</th>
             <th>type</th>
             <th>language</th>
           </tr>
           <xsl:for select="records/record/names">
             <tr>
               <td>
                 <xsl:value-of select="name"/>
               </td>
               <td>
                 <xsl:value-of select="type"/>
               </td>
               <td>
                 <xsl:value-of select="language"/>
               </td>
             </tr>
           </xsl:for>
         </table>
       </body>
     </html>
   </xsl:template>
</xsl:stylesheet>
```

1в:

```
<?xml version="1.0" encoding="utf-16"?>
```

<!--Created with Liquid XML Studio 2012 Developer Edition (Trial) 10.1.6.4255 (http://www.liquid-technologies.com)-->
<?xml-stylesheet type='text/xsl' href='transformer.xsl'?>
<records>
<record>
<qualifier>null</qualifier>
<previous>
<recordFromQualifier>null</recordFromQualifier>
<recordToQualifier>null</recordToQualifier>
<document>
<description>null</description>
<date>null</date>
</document>
</previous>
<names>
<name>
<name>Новосибирск</name>
<type>город</type>
<language>ru</language>
<beginDocument>
<uri>uri of document</uri>
<description>Создание объекта Новосибирск</description>
<date>null</date>
<creationDate>null</creationDate>
</beginDocument>
<endDocument>
<description>unknown document</description>
<date>null</date>
<creationDate>null</creationDate>
</endDocument>
</name>
</names>
<locations>
<point>
<beginDocument>
<uri>uri of document</uri>
<description>Новосибирск</description>
<date>null</date>
<creationDate>null</creationDate>
</beginDocument>
<endDocument>
<description>unknown document</description>
<date>null</date>
<creationDate>null</creationDate>
</endDocument>
<latitude>55.03</latitude>
<longitude>82.91</longitude>
</point>

```
        </locations>
        <contains />
        <belongTo />
    </record>
</records>
```

Применение предложенного метода загрузки записей в тезаурус ретроспективного геокодирования позволяет снизить трудоемкость поддержки находящихся в нем данных в актуальном состоянии, а также упрощает его пополнение новыми данными.

ЛИТЕРАТУРА

[1]. Жижимов О.Л., Мазов Н.А. Проблемы географической привязки цифровых объектов в электронных библиотеках // Труды Двенадцатой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2010). – Казань, 13-17 октября 2010 г. С. 207-214.

[2]. 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

[3]. Скачков Д.М., Жижимов О.Л. Об использовании ретроспективного геокодирования для географического поиска в электронных библиотеках // Труды Тринадцатой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2011). – Воронеж, 19-22 октября 2011 г. С. 30-37.

[4]. Бенкен Е. «PHP, MySQL, XML Программирование для интернета» 3-е издание. – Санкт-Петербург «БХВ-Петербург» 2011г. С 231-235.

[5]. Геокодирование - Службы API Карт Google. https://developers.google.com/maps/

[6]. Liquid technologies. http://www.liquid-technologies.com