

# **НОВЫЕ МЕТОДЫ И ПОДХОДЫ В КАРТОГРАФИИ И ГЕОИНФОРМАТИКЕ**

## **NEW METHODS AND APPROACHES IN CARTOGRAPHY AND GEOINFORMATICS**

УДК: 528.94

DOI: 10.35595/2414-9179-2021-2-27-114-127

Л.А. Обухов<sup>1</sup>, Е.А. Паниди<sup>2</sup>

### **О КОНТРОЛЕ КОРРЕКТНОСТИ ПРИ ГЕОКОДИРОВАНИИ ПОЧТОВЫХ АДРЕСОВ**

#### **АННОТАЦИЯ**

В статье рассмотрено содержание и результаты разработки методики геокодирования почтовых адресов. Задача геокодирования рассмотрена на примере исследования, посвящённого изучению пространственного распределения и динамики случаев инфицирования туберкулёзом и сопутствующими заболеваниями. Исследование выполняется в масштабе крупного города, на примере Санкт-Петербурга (Россия).

Предложенная методика основана на дополнении классической схемы геокодирования, предполагающей прямое связывание адресных данных, представленных в составе исходных данных, с адресными данными в справочнике (в базе пространственных данных). Дополнение состоит в использовании эталонного реестра почтовых адресов, в качестве которого задействуется адресная база данных, создаваемая государственными органами. Записи исходных данных связываются с записями эталонного реестра почтовых адресов, а записи реестра, в свою очередь, связываются с записями справочника, используемого для геокодирования (адресами, записанными в атрибуты объектов базы пространственных данных).

Указанный подход позволяет обеспечить контроль корректности структуры и содержания адресных данных, используемых при геокодировании, а также привести адресные данные к унифицированной форме, официально закреплённой на государственном уровне и используемой в официальных документах и информационных системах.

Предложенная методика реализована на примере системы почтовых адресов, используемой в Российской Федерации, однако, она также может быть использована при работе с адресными системами других государств. В таком случае, для реализации методики необходимо иметь официальный реестр почтовых адресов, представленный в структурированном виде (желательно, в форме базы данных). Методика может быть использована как при геокодировании данных медицинской статистики, так и при решении задач в других предметных областях.

**КЛЮЧЕВЫЕ СЛОВА:** QGIS, ФИАС, Nominatim, KLADR-API, OSM

<sup>1</sup> Санкт-Петербургский государственный университет, институт наук о Земле, кафедра картографии и геоинформатики, Санкт-Петербург, Россия; *e-mail*: [obuhov.lev@mail.ru](mailto:obuhov.lev@mail.ru), [st068972@student.spbu.ru](mailto:st068972@student.spbu.ru)

<sup>2</sup> Санкт-Петербургский государственный университет, институт наук о Земле, кафедра картографии и геоинформатики, Санкт-Петербург, Россия; *e-mail*: [panidi@ya.ru](mailto:panidi@ya.ru), [e.panidi@spbu.ru](mailto:e.panidi@spbu.ru)

Lev Obuhov<sup>1</sup>, Evgeny Panidi<sup>2</sup>

## TOWARD CORRECTNESS CONTROL OF POSTAL ADDRESSES GEOCODING

### ABSTRACT

The paper discusses content and results of the methodology elaborated for geocoding of postal addresses. The geocoding issue is considered on the example of study devoted to the exploration of the spatial distribution and dynamics tuberculosis and concomitant diseases infection cases. The study is carried out on a large city scale. The example of St. Petersburg city (Russia) is used.

Proposed methodology is based upon the extending of the classical geocoding scheme that assumes direct linking of the address data presented as a part of initial dataset with the address data presented in the reference dataset (in the geospatial database). The extension consists in the use of a middle reference-standard register of postal addresses. An address database developed by official agencies is used as a reference-standard register. Initial data records are linked with the records of the reference-standard register by postal addresses, and the register records, in turn, are linked with the records of the reference dataset used for geocoding (with the addresses recorded in the attributes of the geospatial database objects).

This approach allows to provide control of structure and content correctness for the address data used for geocoding, as well as to convert address data into a unified form accepted officially at the state level and used in official documents and information systems.

The methodology is implemented on the example of the postal address system used in the Russia. However, it can also be used when operating with the address system of any other states. In such a case, it is necessary to have an official register of postal addresses to implement the methodology. The register have to be presented in a structured form (preferably in the form of a database). The methodology can be used both for medical statistics data geocoding, and for geocoding of other domain data.

**KEYWORDS:** QGIS, FIAS, Nominatim, KLADR-API, OSM

### ВВЕДЕНИЕ

В настоящей статье рассмотрен подход к решению задачи геокодирования данных медицинской статистики при выполнении исследований в области медицинской географии и картографии, предполагающих привлечение первичных медицинских данных, извлекаемых непосредственно из медицинских карт пациентов. В медицинских картах, оформляемых при регистрации пациентов, имеется информация об адресах места жительства. Используя их возможно сформировать базу пространственных данных, которая в дальнейшем может быть использована для выполнения аналитических выкладок средствами географических информационных систем (ГИС) и изучения характера и тенденций территориального распределения различных заболеваний. Для минимизации количества ошибок геокодирования, необходимо обеспечить контроль корректности адресов, извлекаемых из медицинских карт.

Необходимо отметить, что термин «геокодирование» имеет несколько толкований. В настоящей статье идёт речь о так называемом адресном геокодировании, широко применяемом при наполнении географических информационных систем данными [*Борис и*

<sup>1</sup> Saint Petersburg State University, Institute of Earth Sciences, Department of Cartography and Geoinformatics, St. Petersburg, Russia, *e-mail:* obuhov.lev@mail.ru, st068972@student.spbu.ru

<sup>2</sup> Saint Petersburg State University, Institute of Earth Sciences, Department of Cartography and Geoinformatics, St. Petersburg, Russia, *e-mail:* panidi@ya.ru, e.panidi@spbu.ru

др., 2014; Фарафонов, 2014; Wheeler, 2020] и предполагающем автоматическое формирование геометрии (объекта карты) на основе описания местоположения объекта (представленного в форме почтового адреса). При этом принято выделять прямое и обратное геокодирование. Прямое предполагает извлечение из некоторого массива исходных данных описания положения объекта с последующим формированием его геометрической модели в базе пространственных данных и присвоением ей в качестве атрибутов иных данных об объекте, которые могут быть извлечены из массива исходных данных. Обратное геокодирование, в свою очередь, подразумевает получение описания положения (адреса) объекта по его координатам. Геокодирование используется для пространственного поиска, например, по названиям улиц или городов, нахождения расположения домов по их адресу, поиска расположения различных организаций. Использование геокодирования данных медицинской статистики также применяется достаточно часто и имеет свои особенности [Rashidian et al., 2017, 2018; Rizwan, 2018], в том числе связанные с необходимостью обеспечить обезличенность данных [Harris, Delcher, 2019].

С точки зрения алгоритмизации задача геокодирования заключается в поиске наилучшего совпадения описания местоположения с атрибутами в адресной базе имеющейся в составе вспомогательной базы пространственных данных, справочнике, на основе которой и выполняется геокодирование. Поиск строгого совпадения использовать затруднительно, так как синтаксис и грамматика описания местоположения приемлемые для человека и информационной системы, в общем случае, существенно различаются. Например, записи «ул. Ботаническая» и «улица Ботаническая» в понимании человека являются идентичными, однако для программной системы это два разных адреса. В связи с этим, поиск совпадений описания адресов в массиве исходных данных и в атрибутах справочника, как правило, допускает нестрогое соответствие, что в свою очередь, ведёт к появлению ненулевой вероятности возникновения ошибок геокодирования. Кроме того, качество и точность данных, хранимых в справочниках, также влияет на возможность возникновения ошибок геокодирования и, собственно, на возможность выполнения геокодирования.

## МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

В настоящей работе задача геокодирования рассматривается в контексте исследования выполняемого в 2019–2021 годах и посвящённого картографированию пространственного распределения и динамики заболеваемости туберкулёзом и сопутствующими инфекционными заболеваниями в масштабе города [Kuznetsov et al., 2020]. В качестве исходных данных в исследовании использованы обезличенные регистрационные карточки, заполняемые в районных туберкулёзных диспансерах при выявлении и регистрации случаев инфицирования туберкулёзом. Исследование выполняется на примере Санкт-Петербурга. Исходные данные собраны и предоставлены Санкт-Петербургским НИИ Фтизиопульмонологии<sup>1</sup>. Особенностью исходных данных явилось то, что значительная их часть изначально представлена в аналоговой форме (заполненные вручную бумажные медицинские формы), что существенно повысило возможность появления ошибок в данных (неверное написание названий улиц, ошибочные или отсутствующие номера или корпуса домов, и др.). К моменту подготовки данного текста всего обработано несколько тысяч таких форм. Первоначально данные заносятся в электронную таблицу (рис. 1), которая в дальнейшем подвергается автоматизированному переструктурированию и геокодированию.

<sup>1</sup> <https://spbniif.ru>

ID	район спб	тип	название	ном.	ном.	год рожд	пол	дата выявл	дата снятия	бк+	млу	тб+	тб+
				дома	корп								
1	петроград	ул	гатчинская	31/33		06 12 1966	ж	27 11 2012	18 04 2014				
2	петроград	пр	кронверский	27		02 01 1960	м	16 11 2012	06 08 2013				
3	петроград	ул	саблинская	13/15		11 07 1954	м	01 01 2013	08 05 2014	бк+			
4	петроград	ул	куйбышева	10		29 11 1952	ж	10 12 2012	13 09 2013				
5	петроград	ул	пионерская	35		28 07 1963	ж	01 02 2013	24 03 2014	бк+			
6	петроград	пр	добрлюбова	3		25 09 1980	ж	01 02 2013	30 04 2014	бк+			
7	петроград	ул	б.монетная	9		25 07 1980	м	17 01 2013	20 01 2017	бк+			
8	петроград	ул	яблочкова	3		04 05 1970	м	05 02 2013					
9	петроград	ул	куйбышева	36		07 06 1965	м	29 01 2013	11 12 2013				
10	петроград	пер	вяземский	6		16 07 1987	ж	30 11 2012	21 03 2014				
11	петроград	ул	кропоткина	11		05 01 1963	ж	25 12 2012	06 08 2013				
12	петроград	ул	свезжинская	26/28		31 07 1993	м	28 02 2013		бк+	млу		
13	петроград	ул	гатчинская	18		31 08 1977	ж	28 02 2013	06 06 2014				
14	петроград	ул	резная	19/8		17 08 1958	ж	01 04 2013		бк+			
15	петроград	ул	б.зеленина	19		18 03 1986	м	17 01 2013	27 02 2014			в23	
16	петроград	ул	б.пушкарская	39		12 01 1991	м	12 03 2013	17 04 2014				
17	петроград	ул	м.посадская	15		08 10 1983	ж	07 02 2013		бк+	млу	в23	

*Рис. 1. Пример таблицы с исходными данными о зарегистрированных случаях инфицирования*

*Fig. 1. Example of the table containing initial data on registered infection cases*

Почтовый адрес в данном случае разбит на несколько полей, «тип», в котором хранится информация о типе объекта улично-дорожной сети (улица, проспект и т.п.), «название» – собственно наименование улицы, «номер дома» и «номер корпуса» (корпус или литера). В совокупности эти поля позволяют выполнить сопоставление адреса с атрибутами справочного хранилища пространственных данных при выполнении геокодирования.

Вместе с тем, содержание адресных данных в приведённом примере не полностью пригодно для выполнения геокодирования из-за наличия ряда систематических неоднозначностей. В частности, некоторые сокращения толкуются двояко, например, «пр.» может расшифровываться и как «проспект», и как «проезд». Использование же сокращений в наименованиях улиц также может привести к ошибкам геокодирования при наличии адресов со схожими наименованиями. Допустимыми являются только сокращения, принятые на законодательном уровне, например, сокращение «П.С.» (от «Петроградской стороны»). Наконец, далеко не всегда сдвоенные номера домов записываются через знак «/», кроме него может использоваться, к примеру, дефис. Перечисленные систематические неточности неизбежно дополняются грамматическими ошибками в наименованиях улиц, так как массив исходных данных, в описываемом случае, формируется вручную.

Для обеспечения корректного геокодирования необходимо в первую очередь привести запись адресов к унифицированной форме (совместимой с используемой в привлекаемом для геокодирования справочнике). Выявление же и устранение грамматических ошибок целесообразно выполнять уже в процессе геокодирования, формируя списки адресов, для которых в справочнике не были найдены соответствия.

В последние годы устоявшейся практикой является использование для выполнения адресного геокодирования онлайн-сервисов [Степанова, Зайцева, 2015; Pietro, Rinnone, 2017; Silveira et al., 2017; Макаров, 2019; Маннанов, Агшиев, 2020], доступных в глобальной сети, в частности сервисов компаний Яндекс и Google, использующих для геокодирования адресные базы, имеющиеся в составе веб-картографических ресурсов, разрабатываемых и поддерживаемых данными Интернет-компаниями, доступ к которым для выполнения

геокодирования предоставляется на коммерческих условиях, за плату. Альтернативой является веб-сервис Nominatim и соответствующее программное обеспечение<sup>1</sup>, разрабатываемые как свободно распространяемое программное обеспечение с открытым исходным кодом и использующие в качестве справочника данные OpenStreetMap<sup>2</sup> (OSM). Последние также лицензируются как открытые и свободно распространяемые данные. Однако, будучи данными, создаваемыми пользователями, данные OSM, во многих случаях, имеют невысокое качество адресной базы (ошибки и пропуски в адресных данных).

Рассматриваемое исследование не предполагало возможность использования коммерческих систем геокодирования, в связи с чем методика геокодирования была разработана на основе использования Nominatim, а для повышения надёжности геокодирования, контроля его корректности и унификации записи адресов были использованы данные Федеральной информационной адресной системы<sup>3</sup> (ФИАС). ФИАС является государственной информационной системой, предназначенной для формирования, ведения и использования государственного адресного реестра. Базы данных почтовых адресов, формируемые и поддерживаемые в ФИАС, находятся в открытом доступе. Для проверки корректности адреса при геокодировании, в рамках описываемой методики, производится проверка его наличия в базе данных ФИАС. Если совпадение найдено, адрес считается верифицированным.

При автономном использовании ФИАС на настольном компьютере необходимо развернуть базу данных объёмом в несколько гигабайт (файл формата DBF либо XML). Такой вариант организации работы оказывается неудобным, так как предполагает возможность выполнения геокодирования только на подготовленном и оснащённом рабочем месте. В связи с этим, для контроля корректности адресов был применён веб-сервис «ФИАС в облаке»<sup>4</sup> (KLADR-API), являющийся коммерческим, но предоставляющий бесплатный тариф с ограничениями. С использованием данного веб-сервиса данные ФИАС могут быть загружены в автоматическом режиме в формате JSON<sup>5</sup>.

Контроль корректности адреса и приведение его к унифицированной форме производится следующим образом. Первоначально выполняется HTTP-запрос к сервису KLADR-API (Рис. 2) для проверки корректности названия улицы.

<https://kladr-api.ru/api.php?query=декабристов&contentType=street&cityId=7800000000000>

*Рис. 2. Пример HTTP-запроса к веб-сервису KLADR-API для поиска по названию улицы*  
*Fig. 2. Example of the HTTP request to the KLADR-API Web service, street name search*

В этом запросе «декабристов» – наименование улицы, взятое из состава данных геокодируемого адреса, запись «contentType=street» указывает на то, что ответом сервиса на запрос должно быть именно название улицы, а не иные наименования, запись «cityId=7800000000000» указывает, что результаты должны возвращаться только для Санкт-Петербурга.

Формат ответа сервиса представлен на рисунке 3. Данные возвращаются в виде структурированного текстового массива в формате JSON, включающего несколько разделов. В разделе «searchContext» приведены исходные данные, по которым осуществлён поиск, в разделе «result» – нумерованные разделы с найденными результатами. При этом

<sup>1</sup> <https://nominatim.org>

<sup>2</sup> <https://www.openstreetmap.org>

<sup>3</sup> <https://fias.nalog.ru>

<sup>4</sup> <https://kladr-api.ru>

<sup>5</sup> <https://www.json.org/json-en.html>



подраздел «0» в составе раздела «result» является служебным, а сами результаты поиска содержатся в подразделе «1» и последующих. Подраздел «2» и следующие за ним будут присутствовать если найдено более одного соответствия. Например, в представленном случае по наименованию «Декабристов» были найдены улица Декабристов и переулок Декабристов.

```

▼ searchContext:
  contentType: "street"
  cityId: "780000000000"
  query: "декабристов"
▼ result:
  ▼ 0:
    id: "Free"
    name: "Бесплатная версия kladr-api.ru"
    zip: "100000"
    type: ""
    typeShort: ""
    okato: ""
    contentType: "street"
  ▼ 1:
    id: "78000000000035100"
    name: "Декабристов"
    zip: 190121
    type: "Улица"
    typeShort: "ул"
    okato: null
    contentType: "street"
    guid: "2a70149e-f2dd-4147-a2f9-ca54df911365"
    ifnsfl: ""
    ifnsul: ""
    oktmo: ""
    parentGuid: "c2deb16a-0330-4f05-821f-1d09c93331e6"
    cadnum: ""
  ▼ 2:
    id: "78000000000035300"
    name: "Декабристов"
    zip: 199155
    type: "Переулок"
    typeShort: "пер"
    okato: "40263565000"

```

Рис. 3. Пример JSON-ответа на запрос к веб-сервису KLADR-API по наименованию улицы  
Fig. 3. Example of the JSON response on the street-name-based request to the KLADR-API Web service

Далее, из перечня найденных результатов, производится выбор одного из объектов, с учётом типа объекта улично-дорожной сети, указанного в поле «type» геокодируемых данных. Из данных выбранного объекта (подраздела в файле результатов поиска) извлекается значение параметра «id», которое на следующем этапе обработки данных используется для проверки наличия на данной улице номера дома, указанного в составе геокодируемых исходных данных. Для этого выполняется второй HTTP-запрос к сервису KLADR-API (Рис.4).

<https://kladr-api.ru/api.php?query=16&contentType=building&streetId=78000000000035300>

Рис. 4. Пример HTTP-запроса к веб-сервису KLADR-API для поиска по номеру дома  
Fig. 4. Example of the HTTP request to the KLADR-API Web service, house number search

В примере запроса, представленном на рис. 4, «16» это искомый (геокодируемый) номер дома, «contentType=building» – ключ, указывающий на то, что ответ сервиса должен содержать в себе данные о зданиях, «streetId=7800000000035300» – ключ, содержащий значение параметра «id», извлечённого на предыдущем этапе, и указывающего на улицу (объект улично-дорожной сети) для которой необходимо выполнить поиск. В результате выполнения запроса, как и в предыдущем случае, возвращается JSON-файл, содержащий, как правило, один найденный объект. Структура файла (рис. 5) схожа со структурой файла, возвращаемого на первом этапе поиска, описание найденного объекта (здания) представлено в подразделе «1» раздела «result» данного файла.

```

▼ searchContext:
  contentType: "building"
  streetId: "7800000000035300"
  query: "16"
▼ result:
  ▼ 0:
    id: "Free"
    name: "Бесплатная версия kladr-api.ru"
    zip: "100000"
    type: ""
    typeShort: ""
    okato: ""
    contentType: "building"
  ▼ 1:
    id: "780000000003530001"
    name: "16 литера А"
    zip: 199155
    type: "дом"
    typeShort: "д"
    okato: "40263565000"
    contentType: "building"
    guid: "626fcfb9-e215-4a34-ae0a-1fb0fce6b3dd"
    ifnsfl: "7801"
    ifnsul: "7801"
    oktmo: "40311000"
    parentGuid: "9933db3d-8e52-44d8-acbd-c0ccb2096925"
    cadnum: "78:06:0002065:2022"

```

Рис. 5. Пример JSON-ответа на запрос к веб-сервису KLADR-API по номеру дома

Fig. 5. Example of the JSON response on the house-number-based request to the KLADR-API Web service

На данном этапе проводится проверка либо полного совпадения искомого номера дома со значением поля «name» подраздела «1» в возвращённом файле, либо совпадения со значением части поля «name», содержащей в себе число, так как для адресов с литерой «А» обозначение литеры в исходных данных, как правило, не приводится.

При корректном и результативном выполнении двух первых шагов поиска по базе данных ФИАС описание искомого (геокодируемого) адреса может быть сформировано и представлено в структурированном и унифицированном, соответственно представлению в ФИАС, виде. В таком виде описание адреса и переносится в дальнейшем в базу пространственных данных.

Следующим этапом обработки исходных данных является собственно операция геокодирования (получения координат объекта и сохранение их в базе пространственных данных совместно со значениями геокодируемых полей исходных данных). Получение координатной информации выполняется при помощи HTTP-запроса к веб-сервису Nominatim<sup>1</sup> (рис. 6). В данном запросе «self.building\_number», «self.type\_» и «self.name» – это, соответственно, ключи, передающие значения номера дома, типа улицы и наименования улицы, полученные в результате поиска и проверки с использованием ФИАС. Также в составе запроса указывается наименование города и тип данных ответа сервиса (в данном случае JSON-файл).

```
https://nominatim.openstreetmap.org/?street='+self.building_number+'/'+'self.type_'+'+self.name+'&city=Санкт-Петербург&format=geojson
```

Рис. 6. Пример HTTP-запроса к веб-сервису Nominatim  
Fig. 6. Example of the HTTP request to the Nominatim Web service

Пример JSON-ответа на запрос приведён на рис. 7. Координатные и атрибутивные данные, извлечённые из OSM, содержатся в нумерованных подразделах раздела «features» возвращаемого файла. С точки зрения геокодирования основными являются значения в полях подраздела «coordinates», где поле «0» содержит значение долготы, а поле «1» – значение широты центроида найденной по адресу геометрии. Для указания координат используется система координат WGS-84. По извлечённым из данного файла координатам создаётся вспомогательный точечный векторный объект в базе пространственных данных, которому и присваиваются все геокодируемые атрибуты, включая поля почтового адреса, унифицированного на предыдущих этапах обработки данных (рис. 8).

```

type: "FeatureCollection"
  licence: "Data © OpenStreetMap contributors, ODbL 1.0. https://osm.org/copyright"
  features:
    0:
      type: "Feature"
      properties:
        place_id: 214415462
        osm_type: "way"
        osm_id: 605372866
        display_name: "Институт наук о Земле СПбГУ, 16, переулок Декабристов, округ Остров Декабристов, Санкт-Петербург"
        place_rank: 30
        category: "amenity"
        type: "university"
        importance: 0.411
        icon: "https://nominatim.openstreetmap.org/ui/mapicons//education_university.p.20.png"
      bbox:
        0: 30.2473971
        1: 59.9526252
        2: 30.2484687
        3: 59.9534746
      geometry:
        type: "Point"
        coordinates:
          0: 30.247647477171984
          1: 59.9530699

```

Рис. 7. Пример JSON-ответа на запрос к веб-сервису Nominatim  
Fig. 7. Example of the JSON response on the request to the Nominatim Web service

<sup>1</sup> <https://nominatim.openstreetmap.org>

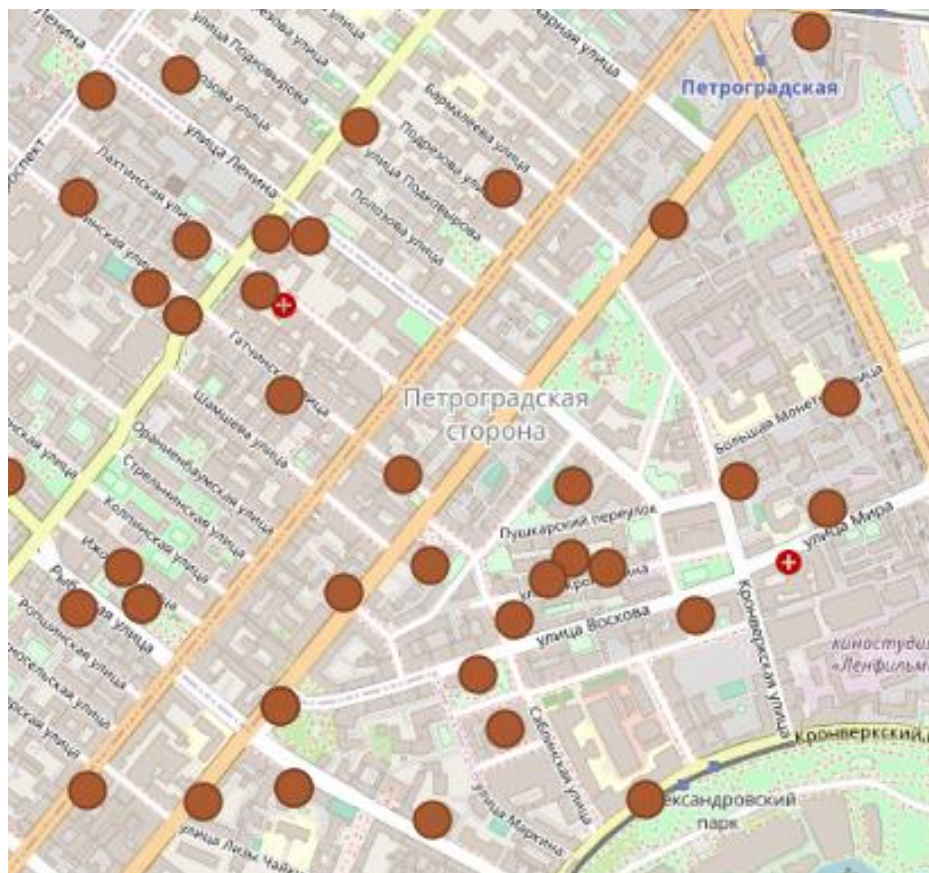


	Адрес	Тип адреса	Номер дома	id пациента	Дата выявления
1	Гатчинская	улица	22 корпус ...	101	2015

*Рис. 8. Отображение геокодированных исходных данных в атрибутивной таблице векторного слоя в QGIS*

*Fig. 8. Representation of the geocoded initial data in attribute table of the vector layer in QGIS*

На заключительном этапе геокодирования из массива данных OSM, путём выполнения автоматической выборки по расположению, на основе координат центроида, полученного на предыдущих шагах, извлекается полигональная геометрия, описывающая положение контура здания. Эта геометрия сохраняется во второй векторный (полигональный) слой, с присвоением ей атрибутов аналогичных атрибутам ранее созданного точечного объекта. В результате, геокодированные данные состоят из двух слоёв, точечного и полигонального. Геокодированные атрибуты (исходные данные) оказываются связанными с точечной (центроидом) и полигональной геометрией. Окончательное структурирование данных, предполагающее или не предполагающее (в случае использования реляционной модели данных) дублирование геокодированных атрибутов для двух типов геометрии (рис. 9, 10) геокодированного объекта, может быть выполнено в соответствии с контекстом решаемой задачи.



*Рис. 9. Отображение слоя геокодированных центроидов зданий в QGIS*

*Fig. 9. Representation of the geocoded layer of centroids in QGIS*



*Рис. 10. Отображение слоя геокодированных полигонов зданий в QGIS*  
*Fig. 10. Representation of the geocoded layer of building polygons in QGIS*

При разработке модуля были использованы следующие программные библиотеки:

- 1) requests<sup>1</sup> (для выполнения HTTP-запросов и приёма ответов от веб-сервисов);
- 2) geojson<sup>2</sup> (для создания, чтения и редактирования файлов формата GeoJSON);
- 3) processing<sup>3</sup> (для автоматизированного применения инструментов обработки данных, доступных в QGIS, в частности для выполнения пространственных запросов);
- 4) qgis.PyQt<sup>4</sup> (для реализации графического интерфейса модуля);
- 5) qgis.core<sup>5</sup> (для автоматизированного создания и добавления на карту векторных слоёв в QGIS).

## РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

Для выполнения тестирования описанной методики геокодирования был разработан модуль QGIS, способный открывать отдельное всплывающее окно (форму) для ввода геокодируемых данных в контексте интерфейса (из интерфейса) QGIS (рис. 11). Модуль разработан с использованием стандартных (для QGIS) средств QGIS Plugin Builder<sup>6</sup>. После ввода исходных геокодируемых данных в поля формы, в результате автоматического выполнения запросов к необходимым веб-сервисам, модуль сохраняет данные в файл в формате GeoJSON<sup>7</sup>, который может быть отображён в QGIS в виде слоя карты, содержащего геокодированные геометрию и атрибуты.

<sup>1</sup> <https://github.com/psf/requests>

<sup>2</sup> <https://github.com/jazzband/geojson>

<sup>3</sup> <https://github.com/qgis/QGIS/tree/master/python/plugins/processing>

<sup>4</sup> <https://github.com/qgis/QGIS/tree/master/python/PyQt>

<sup>5</sup> <https://github.com/qgis/QGIS/tree/master/python/core>

<sup>6</sup> <https://github.com/g-sherman/Qgis-Plugin-Builder>

<sup>7</sup> <https://geojson.org>

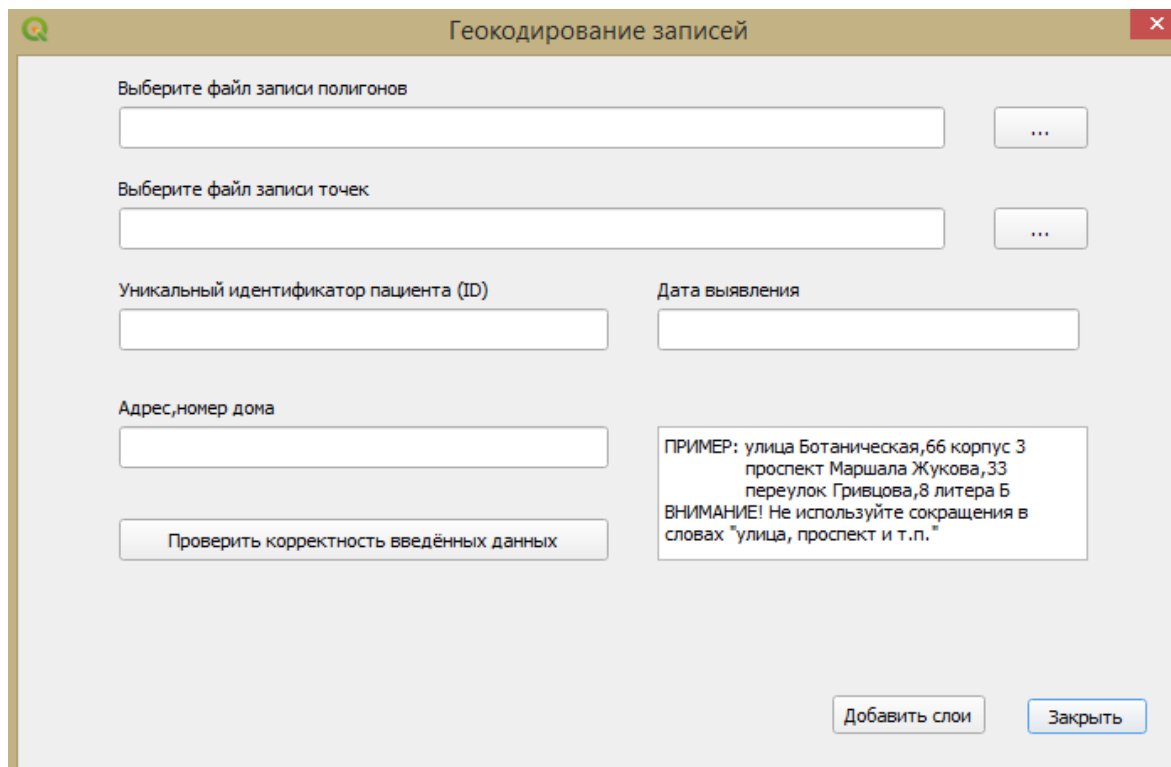


Рис. 11. Интерфейс модуля QGIS, разработанного для тестирования геокодирования  
 Fig. 11. Interface of the QGIS module, developed for the geocoding testing

Предложенная методика геокодирования данных медицинской статистики была апробирована при геокодировании используемых в рамках исследования данных по заболеваемости туберкулёзом. Результаты первичного тестирования продемонстрировали 95 % результативность при проверке и унификации адресов и 87 % результативность при формировании геометрии объектов. Данное различие вызвано ошибками и неполнотой данных OSM. 5 % данных, в свою очередь, были признаны некорректными в связи с ошибками в адресах, приведённых в составе исходных данных. Следует отметить, что при тестировании применялся последовательный ручной ввод геокодируемых данных и ручной контроль сопоставления данных, процедуры выполнялись в автоматизированном, а не полностью автоматическом режиме. Автоматическое геокодирование, ожидаемо, должно дать худшие результаты.

В ходе выполнения обработки данных были выявлены следующие виды ошибок:

1) Ошибки записи данных в медицинские карточки при регистрации случаев инфицирования – в большинстве случаев неуместное использование знака «/» в номерах домов, а также грамматические ошибки в наименованиях улиц и ошибки в указании типа объекта улично-дорожной сети (например, улица вместо проспекта);

2) Ошибки ФИАС. Госреестр адресов, хоть и в минимальном количестве, но всё же содержит ошибки, причиной появления которых, вероятнее всего, является человеческий фактор. Так, адреса по улице Куйбышева (Санкт-Петербург) при первоначальном геокодировании были полностью признаны некорректными, из-за ошибочного дублирования записей в ФИАС;

3) Ошибки получения координат и ошибки связывания атрибутов с полигональной геометрией. В OSM присутствуют ошибки, связанные, в основном, с присвоением домам номеров (например, присвоение дому номера «17» при официальном номере «15–17»).



## ВЫВОДЫ

В сущности, разработанная и протестированная в данном исследовании методика предполагает использование в дополнение к справочнику (адресной базе пространственных данных) верифицированного (официального) адресного реестра, созданного и поддерживаемого официальными службами или организациями, официально наделёнными правом таких работ на уровне государства. В рамках методики, указанный реестр используется как эталон адресов, а геокодируемые данные связываются с данными справочника не напрямую, а через реестр. Фактически записи геокодируемых данных и справочника связываются с эталонным реестром, из которого в итоге и извлекается верифицированный почтовый адрес объекта. В свою очередь, геометрия объекта извлекается из используемого справочника, а тематические (геокодируемые) атрибуты из массива исходных геокодируемых данных.

Описанная методика может быть применена для оптимизации процессов и улучшения качества ведения медицинских баз пространственных данных при выполнении медико-географических и медико-картографических исследований. Она позволяет обеспечить контроль формы и содержания адресных данных, а также приведения их к унифицированному виду. Кроме того, использованный подход может быть применён при формировании медицинских баз данных для контроля сохраняемых в базе адресных данных.

## СПИСОК ЛИТЕРАТУРЫ

1. *Борис Ф.Н., Махт В.А., Борис Е.А.* Решение задач массовой оценки недвижимости с применением метода геокодирования. Омский научный вестник, 2014. № 2(130). С. 214–216.
2. *Макаров О.С.* Обзор online-сервисов по формализации адресов. XLVII Огарёвские чтения, Материалы научной конференции, 2019. С. 352–357.
3. *Маннанов А.А., Агишев Т.Х.* Геокодирование областей средствами Яндекс карт и поисковой системы данных Openstreetmap Nominatim. Теория и практика модернизации научной деятельности в условиях цифровизации, Сборник статей международной научно-практической конференции, 2020. С. 20–23.
4. *Степанова Л.А., Зайцева Е.Н.* Геокодирование объектов в Quantum GIS с использованием базы данных Яндекс. Программные продукты и системы, 2015. № 3. С. 199–203. DOI: 10.15827/0236-235X.111.199-203.
5. *Фарафонов А.И.* Разработка системы автоматического геокодирования в транспортной компании. Вестник ИЖГТУ имени М.Т. Калашникова, 2015. Т. 18. № 1. С. 103–105.
6. *Harris D.R., Delcher C.* Bench4gis: benchmarking privacy-aware geocoding with open big data. Proceedings – 2019 IEEE International Conference on Big Data, 2019. Article 9006234. P. 4067–4070. DOI:10.1109/BigData47090.2019.9006234.
7. *Kuznetsov I., Panidi E., Kolesnikov A., Kikin P., Korovka V., Galkin V.* GIS-based infectious disease data management on a city scale, case study of St. Petersburg, Russia. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2020. V. XLIII-B3-2020. P. 1463–1467. DOI:10.5194/isprs-archives-XLIII-B3-2020-1463-2020.
8. *Kuznetsov I., Panidi E., Korovka V., Galkin V.* City-scale GIS-based monitoring of infectious disease – contemporary issues, case study of St. Petersburg, Russia. Proceedings of the 6th International Scientific Conference GEOBALCANICA, 2020. P. 791–796. DOI: 10.18509/GBP. 2020.87.

9. *Kuznetsov I., Panidi E., Korovka V., Galkin V., Voronov D.* Web-based representation and management of infectious disease data on a city scale, case study of St. Petersburg, Russia. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2020. V. XLIV-3/W1-2020. P. 87–91. DOI: 10.5194/isprs-archives-XLIV-3-W1-2020-87-2020.
10. *Pietro G.D., Rinnone F.* Online geocoding services: a benchmarking analysis to some European cities. *Proceedings of the 2017 Baltic Geodetic Congress (BGC Geomatics)*, 2017. Article 8071486. P. 273–281. DOI: 10.1109/BGC.Geomatics.2017.12.
11. *Rashidian S., Dong X., Avadhani A., Poddar P., Wang F.* Effective scalable and integrative geocoding for massive address datasets. *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2017. Article 26. P. 1–10. DOI: 10.1145/3139958.3139986.
12. *Rashidian S., Jain S.K., Dong X., Wang F.* EaserGeocoder: integrative geocoding with machine learning. *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2018. P. 572–575. DOI: 10.1145/3274895.3274929.
13. *Rizwan M., Dass S.C., Asirvadam V.S., Gill B.S., Sulaiman L.H.* DenMap: a Dengue surveillance system for Malaysia. *Journal of Physics: Conference Series*, 2018. V. 1123. Issue 1. Article 012045. DOI: 10.1088/1742-6596/1123/1/012045.
14. *Silveira I.H.D., Oliveira B.F.A., Junger W.L.* Use of Google Maps for geocoding data from the Mortality Information System in Rio de Janeiro municipality, Brazil, 2010-2012. *Epidemiologia e servicos de saude : revista do Sistema Unico de Saude do Brasil*, 2017. V. 26. Issue 4. P. 881–886. DOI: 10.5123/S1679-49742017000400018.
15. *Wheeler A.P., Gerell M., Yoo Y.* Testing the spatial accuracy of address-based geocoding for gunshot locations. *Professional Geographer*, 2020. V. 72. Issue 3. P. 398–410. DOI: 10.1080/00330124.2020.1730195.

## REFERENCES

1. *Boris F.N., Makht V.A., Boris E.A.* The solution of problems of mass appraisal of real estate with application of the method of geocoding. *Omsk Scientific Bulletin*, 2014. V. 2 (130). P. 214–216 (in Russian).
2. *Harris D.R., Delcher C.* Bench4gis: benchmarking privacy-aware geocoding with open big data. *Proceedings – 2019 IEEE International Conference on Big Data*, 2019. Article 9006234. P. 4067–4070. DOI:10.1109/BigData47090.2019.9006234.
3. *Farafonov A.I.* Development of automatic geocoding in transportation company. *Vestnik IzGTU imeni M.T. Kalasnikova*, 2015. V. 18. Issue 1. P. 103–105 (in Russian).
4. *Kuznetsov I., Panidi E., Kolesnikov A., Kikin P., Korovka V., Galkin V.* GIS-based infectious disease data management on a city scale, case study of St. Petersburg, Russia. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2020. V. XLIII-B3-2020. P. 1463–1467. DOI: 10.5194/isprs-archives-XLIII-B3-2020-1463-2020.
5. *Kuznetsov I., Panidi E., Korovka V., Galkin V.* City-scale GIS-based monitoring of infectious disease – contemporary issues, case study of St. Petersburg, Russia. *Proceedings of the 6th International Scientific Conference GEOBALCANICA*, 2020. P. 791–796. DOI: 10.18509/GBP.2020.87.
6. *Kuznetsov I., Panidi E., Korovka V., Galkin V., Voronov D.* Web-based representation and management of infectious disease data on a city scale, case study of St. Petersburg, Russia. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information*

- Sciences, 2020. V. XLIV-3/W1-2020. P. 87–91. DOI: 10.5194/isprs-archives-XLIV-3-W1-2020-87-2020
7. *Makarov O.S.* Overview of online services for addresses formalizing. XLVII Ogarev Readings, Proceedings of the Scientific Conference, 2019. P. 352–357 (in Russian).
  8. *Mannanov A.A., Agishev T.H.* Geocoding of areas using Yandex Maps and the Openstreetmap Nominatim data search engine. Theory and Practice of Modernization of the Scientific Activity in the Conditions of Digitalization, Proceedings of the International Scientific and Practical Conference, 2020. P. 20–23 (in Russian).
  9. *Pietro G.D., Rinnone F.* Online geocoding services: a benchmarking analysis to some European cities. Proceedings of the 2017 Baltic Geodetic Congress (BGC Geomatics), 2017. Article 8071486. P. 273–281. DOI: 10.1109/BGC.Geomatics.2017.12.
  10. *Rashidian S., Dong X., Avadhani A., Poddar P., Wang F.* Effective scalable and integrative geocoding for massive address datasets. Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2017. Article 26. P. 1–10. DOI: 10.1145/3139958.3139986.
  11. *Rashidian S., Jain S.K., Dong X., Wang F.* EaserGeocoder: integrative geocoding with machine learning. Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2018. P. 572–575. DOI: 10.1145/3274895.3274929.
  12. *Rizwan M., Dass S.C., Asirvadam V.S., Gill B.S., Sulaiman L.H.* DenMap: a Dengue surveillance system for Malaysia. Journal of Physics: Conference Series, 2018. V. 1123. Issue 1. Article 012045. DOI: 10.1088/1742-6596/1123/1/012045.
  13. *Silveira I.H.D., Oliveira B.F.A., Junger W.L.* Use of Google Maps for geocoding data from the Mortality Information System in Rio de Janeiro municipality, Brazil, 2010-2012. Epidemiologia e servicos de saude : revista do Sistema Unico de Saude do Brasil, 2017. V. 26. Issue 4. P. 881–886. DOI:10.5123/S1679-49742017000400018.
  14. *Stepanova L.A., Zaitseva E.N.* Geocoding sites in Quantum GIS with the help of Yandex databases. Software & Systems, 2015. V. 3. P. 199–203 (in Russian). DOI: 10.15827/0236-235X.111.199-203.
  15. *Wheeler A.P., Gerell M., Yoo Y.* Testing the spatial accuracy of address-based geocoding for gunshot locations. Professional Geographer, 2020. V. 72. Issue 3. P. 398–410. DOI: 10.1080/00330124.2020.1730195.
-