МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ГЕОДЕЗИИ И КАРТОГРАФИИ (МИИГАиК)

Лаборатория открытых данных
(ЛОД)

МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ

«Формирование набора открытых данных»

для студентов по направлению подготовки «Прикладная информатика»

Москва, 2023 г.

СОДЕРЖАНИЕ

Введение

Методические рекомендации «Формирование набора открытых данных» предназначены для практикантов с целью получения навыков по созданию, сбору, обработки, оценки и подготовки к публикации наборов открытых данных.

В данных методических рекомендациях представлен следующий тематический план:

1. Концептуальное проектирование набора данных
2. Определение потенциальной востребованности набора данных
3. Создание набора данных с помощью программных средств
4. Предварительная обработка набора данных
5. Оценка качества набора данных
6. Подготовка набора данных к публикации

При выполнении плана рекомендуется в заключение отчета представить сведения о затраченном времени на каждый вид работ. Стоит обратить внимание, что данная рекомендация не обязывает ее выполнять. Время выполнения работ необходимо указывать максимально точно. Пример формы для записи времени выполнения работ представлен в приложении В.

Концептуальное проектирование набора данных

Данный этап предназначен для тех, у кого еще нет сформированного набора данных, который можно выгрузить из базы данных или перенести его из одной информационной системы в другую и он не хранится ни на одном из возможных электронных носителей.

Чтобы определить содержание планируемого набора данных необходимо выполнить задачи следующих блоков:

Блок 1	**Предварительный этап определения содержания набора данных**
Задачи	1. Ознакомиться с существующими наборами открытых данных на официальных сайтах публикаторов открытых данных (Приложение А, таб. 1). 2. Среди предложенных тематик (Приложение А, таб. 2) выбрать ту, которая связана с вашей деятельностью или в области которой вы имеете знания. Если есть затруднения с выбором тематики, перейдите к пункту 3. 3. Проанализировать в каких сферах (категориях) больше всего публикуют наборы открытых данных.
Блок 2	**Определение содержания набора данных**
Задачи	1. Описать какие атрибутивные характеристики должен иметь набор данных. *Пример:* «адрес», «услуги, предоставляемые в летний/зимний период», «количество мест» и т.п. 2. Определить какой тип данных должен иметь каждый из атрибутов набора данных. 3. Дать название набору данных. Название должно точно отражать его предметную область.

Блок 3	**Контроль содержания набора данных**
Задачи	1. Ознакомиться с нормативно-правовой документацией, регулирующих публикацию данных (Приложение А, таб. 3). 2. Определить не попадает ли содержание набора данных под какой-либо из следующих критериев: *a. <u>категория доступа к информации</u>* Подразделяет информацию на различные категории в зависимости от типа доступа к ней и закладывает основы обращения различных лиц с такой информацией. Ограничение доступа к информации производится на основании федеральных законов, при этом соблюдение конфиденциальности такой информации является обязательным *b. <u>порядок распространения информации</u>* • Информация, распространяемая свободно – общеизвестные сведения и иная информация, доступ к которой не ограничен. • Информация, распространение которой ограничено: o реклама; o предвыборная агитация; o нецензурная брань; o сведения и материалы, порочащие репутацию граждан и юридических лиц; o изображения граждан (в том числе несовершеннолетних); o охраняемые результаты интеллектуальной деятельности. • Информация, распространение которой запрещено (см. Федеральный закон от 27.07.2006 N 149-ФЗ (ред. от 24.06.2025) "Об информации, информационных технологиях и о защите информации")/Статья 10 пункт 6 подпункты а-к.

Блок 3	Контроль содержания набора данных
Задачи	• Информация, распространение которой специально маркируется. Например, к данной категории относится информация, созданная некоммерческими организациями, внесенными в реестр Министерства юстиции - «реестр ОРГАНИЗАЦИЯ, выполняющих функции иностранного агента». *c. <u>гражданско-правовой режим</u>* • Неохраняемые объекты o Неохраноспособные объекты. К ним относятся товарные знаки, которые вошли во всеобщее употребление для обозначения товаров определенного вида. Также объектами, не попадающими под его охрану, являются результаты интеллектуальной деятельности, противоречащие принципам гуманности и морали, общественным интересам. Не получают охрану в качестве изобретений идеи, научные теории и математические выводы, открытия. o Объекты, охрана которых прекратилась в связи с истечением установленных сроков или по воле правообладателя. • Охраняемые объекты o Объекты авторского права (произведения). o Объекты смежных прав (исполнения, фонограммы, сообщение в эфир или по кабелю теле- и радиопередач; содержание баз данных; произведения в рамках публикаторского права). o Ноу-хау (секреты производства, имеющие действительную или потенциальную коммерческую ценность вследствие неизвестности их третьим лицам).

Данный этап необходимо оформить в соответствии с примером, указанным в приложении Б. После оформления можно перейти к этапу «Определение потенциальной востребованности набора открытых данных».

Определение потенциальной востребованности набора данных

1. *Определение популярности тематики*

Необходимо выбрать один из способов определения.

Способ прямого определения: Проведение опроса и/или анкетирования.

Пример содержания анкеты

По каким направлениям (тематикам) необходимо обеспечить открытость данных? (Оценить по шкале от 1 до 10, учитывая, что 10 – наиболее важно обеспечить открытость).

- Компании;
- Преступность и правосудие;
- Наблюдение за планетой;
- Образование;
- Энергетика и окружающая среда;
- Финансовые и контрактные вопросы;
- Геопространственные данные;
- Глобальное развитие;
- Подотчетность правительственного аппарата и демократия;
- Здравоохранение;
- Наука и исследования;
- Статистика;
- Социальная мобильность и благосостояние;
- Транспорт и инфраструктура.

Число анкет или опрошенных респондентов должно быть не менее 10. Желательно сопроводить расчеты анализом проведенного анкетирования/опроса (возрастной состав респондентов, квалификация/сфера деятельности), чтобы объяснить полученные результаты.

Определите с помощью формулы популярность тематики, в которой определен набор данных:

$$К_{\text{тематики}} = \frac{\text{Сумма баллов, полученных тематикой}}{\text{Общая сумма баллов для всех тематик}}.$$

Способ косвенного определения: Определение наиболее востребованной тематики в источниках открытых данных с системой классификации.

При наличии у источника открытых данных системы классификации необходимо определить какие классы относятся к тематике формируемого набора данных и к тематикам, представленным в таблице 2.

Посчитать количество опубликованных наборов открытых данных по каждой тематике и рассчитать для конкретной тематики набора данных коэффициент k_1:

$$k_{i1} = \frac{n_i}{\sum n},$$

где i – порядковый номер тематики, n_i – количество наборов открытых данных конкретной тематики, $\sum n$ – сумма всех наборов по каждой тематике.

Посчитать среднее число скачанных наборов данных по каждой тематике и рассчитать для конкретной тематики набора данных коэффициент k_2:

$$k_{i2} = \frac{m_i}{\sum m},$$

где m_i – среднее значение числа скачанных наборов открытых данных конкретной тематики, $\sum m$ – сумма всех средних значений числа скачанных наборов данных по каждой тематике.

Посчитать среднее число просмотренных наборов данных по каждой тематике и рассчитать для конкретной тематики набора данных коэффициент k_3:

$$k_{i3} = \frac{l_i}{\sum l},$$

где l_i – среднее значение числа просмотренных наборов открытых данных конкретной тематики, $\sum l$ – сумма всех средних значений числа просмотренных наборов данных по каждой тематике.

Далее необходимо рассчитать коэффициент популярности тематики, в которой определен набор данных:

$$К_{тематики} = \frac{k_1 + k_2 + k_3}{3}$$

2. *Анализ поисковых запросов*

Используя один из основных поисковых сервисов, осуществляющих открытую статистику запросов, необходимо определить количество запросов по названию (или ключевым словам) формируемого набора (*Пример*: «Дворовые территории» – 24 176 показов в месяц согласно Яндекс.Wordstat). Определить в том же поисковом сервисе количество запросов по названию «открытые данные». Определение проводить за один период (неделю, месяц, год), который не должен превышать одного года с начала проведения анализа.

Рассчитать коэффициент востребованности набора данных по результатам анализа поисковых запросов:

$$К_{востр1} = \frac{v}{w},$$

где v – количество запросов по названию (или ключевым словам) формируемого набора, w – количество запросов по названию «открытые данные».

Если $v \geq w$, то $К_{востр1} = 1$.

Если $К_{востр1} = 0$, то необходимо исключить данный показатель из расчета общего коэффициента потенциальной востребованности набора данных.

Рекомендованные поисковые сервисы, предоставляющих статистику запросов: Яндекс.Wordstat.

3. *Востребованность набора данных по результатам опроса/анкетирования*

Необходимо провести опрос или анкетирование любым из способов. Рекомендуется использовать механизмы социальных сетей или формы для опроса.

Число заполненных анкет или опрошенных респондентов должно быть не менее 10. Используйте 5-балльную шкалу для оценивания набора.

По результатам опроса или анкетирования необходимо произвести расчет коэффициента востребованности для данного типа исследования:

$$К_{востр2} = \frac{t}{p},$$

где p – сумма всех баллов, выставленных набору данных, t – сумма баллов по каждой из заполненных анкет с 4 и 5 баллами (максимальный балл 4 и 5 × количество опросов/анкет).

4. *Расчет общего коэффициента потенциальной востребованности набора данных*

Для расчета общего коэффициента необходимо воспользоваться следующей формулой:

$$К_{востр\ общ} = \frac{К_{востр1} + К_{востр2} + К_{тематики}}{3} \times 100\%$$

Таблица 1. Группы распределения классов потенциальной востребованности набора данных

Группа	**Признак отнесения к группе**	**Класс востребованности**
Группа A	$К_{востр\ общ} \geq 67\%$	высокая потенциальная востребованность набора данных
Группа B	$К_{востр\ общ} \geq 33\%$ и $< 67\%$	средняя потенциальная востребованность набора данных
Группа C	$К_{востр\ общ} \geq 1\%$ и $< 33\%$	низкая потенциальная востребованность набора данных

В случае если набор данных относится к группе B, C или имеет коэффициент востребованности меньше 1, необходимо предложить возможные варианты повышения класса востребованности.

Создание набора данных с помощью программных средств

При создании набора данных можно использовать один из предложенных способов: использование средств записи данных, использование мобильных средств для сбора пространственных данных и автоматизированный сбор данных. Другие, не представленные способы, могут быть использованы только по согласованию.

1. *Программные средства записи данных*

Программные средства записи данных предназначены для ввода информации в интерактивном режиме. Среди них стоит выделить те, которые представляют данные в формате удобном для обмена: текстовые редакторы, табличные процессоры и информационные системы, позволяющие вводить данные в табличном формате (СУБД).

1.1. Табличный процессор

Для ввода данных можно использовать любой табличный процессор, который поддерживает выгрузку данных в форматах XML, JSON, CSV.

Данный способ предназначен для ввода широкого тематического круга данных. Однако некоторые специфические моменты лучше решать с использованием других способов.

Особенности использования табличных процессоров для создания набора данных:

1. Создание данных без пространственной специфики;
2. Создание новых данных;
3. Создание данных, не требующих сложных связей.

Первый и второй пункты предполагают использования других способов, описанных после первого подраздела.

Рекомендуемые табличные процессоры представлены в Приложении А в таблице 4.

1.2. Системы управления базами данных

Создание данных в базе данных предполагает основательный подход к правильной организации ее структуры.

Так как данные методические рекомендации предполагают формирование одного набора данных, то рекомендуется использовать табличные процессоры. Если есть навыки работы с определенной СУБД, то по согласованию, возможно, реализовать проектирование БД.

Особенности характерные табличным процессорам для создания набора данных применимы и для данного средства.

Чтобы данными можно было пользоваться, необходимы расшифровки и пояснения используемых сокращений, обозначений, единиц измерения и т.п. Это особенно актуально в случаях, когда публикуются таблицы или иные структурированные данные, в которых часто бывают сокращенными названия полей (столбцов) или приведены только величины без их размерности. Например, если столбец называется «сумма», то требуется расшифровка, в которой бы указывалось, в чем эта сумма измеряется. Справочники лучше всего делать в виде отдельных файлов — документов или таблиц.

В зависимости от выбранного средства записи данных необходимо:

- При использовании табличного процессора – представить снимок области, ограничивающей содержание таблицы.
- При использовании БД – представить ER-диаграмму базы данных и снимок области, ограничивающей содержание таблицы, ее информационное наполнение.

2. *Мобильные средства для сбора пространственных данных*

Данный способ подразумевает использование мобильных геоинформационных систем. Мобильные ГИС возможно использовать в полевых условиях, что позволяет: вводить и редактировать данные, записывать треки, строить маршруты. Главной особенностью является создание данных с пространственной спецификой.

Существует большое количество ГИС, которые отличаются функционалом, платформой и целями применения. Следует выбрать ту ГИС, которая отвечает требованиям и специфике набора данных.

При использовании мобильных ГИС необходимо выполнить:

1. Сбор данных, связанный со спецификой набора данных;
2. Выгрузить данные в формате обмена данными.

Таблица 2. Этапы сбора данных

1.	**Подготовительный этап**
a.	Выбрать исследуемый участок (определить район работ) и изучить местность с использование картографических сервисов
b.	Составить список объектов
c.	Разработать оптимальный маршрут
2.	**Полевой этап**
a.	Сбор первичных данных
b.	Производить фотофиксацию объектов
3.	**Камеральный этап**
a.	Выгрузить данные в одном из следующих форматов: CSV/GeoJSON/GML/GeoPackage/Shapefile
b.	Оценить полноту выполненных работ; Есть ли неисследованные территории и где необходимо уточнить маршрут
c.	Определить требуется ли классификация объектов и формирование новых наборов
4.	**Обработка результатов полевого исследования**
a.	Загрузить данные в ГИС
b.	Использовать инструментарий ГИС для решения задач, связанных со спецификой формируемого набора данных
c.	Выгрузить данные в одном из форматов обмена данными: CSV, JSON, XML, GeoJSON, GML

При использовании мобильных ГИС необходимо указать:

1. Название мобильной ГИС;
2. Объяснение выбора данной ГИС или краткий отзыв о программе;
3. Инструменты, функциональные возможности, использованные при создании набора данных;
4. Снимок экрана мобильного ГИС с картографическим интерфейсом;
5. Снимок экрана мобильного ГИС с атрибутивной таблицей;
6. Снимок экрана таблицы, загруженной в табличном процессоре.

Рекомендуемые мобильные ГИС представлены в Приложении А в таблице 5.

3. Автоматизированный сбор данных (Веб-скрейпинг)

Сбор открытых данных, опубликованных в сети «Интернет», может быть осложнён следующими факторами:

- Данные могут быть «растворены» в структуре общедоступных HTML-страниц, являясь значениями атрибутов специфических узлов HTML-дерева.
- Элементы, представляющие интерес в смысле извлечения информации, могут отображаться только по наступлению некоторого пользовательского события. То есть собираемые данные хранятся на сервере в неизвестном виде и посылаются клиенту только после отправки специфического запроса.
- Количество «информационных узлов» HTML-страниц и самих страниц может быть таким большим, что это сделает невозможным своевременное и полное извлечение открытых данных вручную.
- Доля интересующих данных относительно всей информации, заключённой в HTML-документе может быть небольшой, и простое сохранение исходного кода документа в виде текстового файла не позволит

эффективно применять заключённые в нём данные для решения прикладных задач.

Все вышеупомянутые сложности могут быть решены в том случае, если для сбора данных практикант опирается на помощь специальных программных средств – так называемых веб-скрейперов.

Веб-скрейпинг, или автоматизированный сбор открытых данных в сети интернет, - комплекс мероприятий, направленный на получение открытой информации из сети «интернет», не представленной в привычной, машиночитаемой форме. Веб-скрейпинг как правило подразумевает выполнение практикантом обязательных и последовательных этапов: этап «разведки», этап выбора программных средств сбора, «доводка» алгоритма и постобработка собранных данных в машиночитаемом формате.

Таблица 3. Этапы автоматизированного сбора данных

1.	**Подготовительный этап (этап «разведки»)**
a.	Выбор сайта, содержащего предназначенные для извлечения данные (например, сайта-агрегатора).
b.	Анализ исходного кода сайта средствами браузера с целью выявления его структуры и местонахождения в ней тегов с данными. Определение структуры выходных данных (например, наименования полей).
c.	Заключение выводов о портале на предмет автоматизации извлечения данных.
2.	**Выбор программных средств автоматизации**
a.	Определение программных средств автоматизации, согласующихся с выводами, сделанными по итогам этапа 1.
b.	Формирование программного стека автоматизации (приведение сайта к необходимому состоянию, извлечение данных, сериализация).

Таблица 3. Продолжение

3.	**Разработка и реализация алгоритма.**
a.	Составление блока инструкций приведения сайта в необходимое для разбора состояние (запрос-ответ, эмуляция пользовательских событий, рендер).
b.	Составление блока инструкций навигации по структуре HTML-страницы.
c.	Составление блока инструкций сериализации извлечённых данных в машиночитаемом формате.
4.	**Постобработка собранных данных.**
a.	Проверка сериализованных данных на соответствие спецификации машиночитаемого формата.
b.	Отладка и модификация алгоритма, реализованного на этапе 3.
c.	Публикация синтаксически корректных данных, собранных модифицированным алгоритмом.

Предполагается использование вычислительной среды Python и программных пакетов, импортируемых с помощью пакетного менеджера и находящихся в свободном доступе (requests, BeautifulSoup, Selenium). Возможно использование других вычислительных сред и программных инструментов. По факту выполнения, необходимо:

- Сформулировать вывод об архитектуре портала, который предполагается подвергнуть автоматизированной обработке, на предмет веб-скрейпинга, а также программных средствах. Вывод обосновать.
- Предоставить код программного скрипта на языке Python, выполняющего автоматизированный сбор данных. Если выбрана другая программная среда, предоставить код на соответствующем управляющем языке. Если среда не использует скриптинг для управления собственными функциями, описать алгоритм действий либо на естественном языке, либо с помощью блок-схемы. Описание снабдить иллюстрациями.

- Загрузить результат автоматизированного извлечения данных в виде файла одного из машиночитаемых форматов.

Предварительная обработка набора данных

Данный этап предназначен для тех, у кого имеется сформированный набор данных. В рамках выполнения заданий «Предварительная обработка данных» участники должны сформировать основные компетенции в более общем направлении науки о данных. Процесс работы включает в себя ряд операций, направленных на подготовку набора данных к обработки методами машинного обучения.

Для этого необходимо выполнить задачи следующих блоков:

Блок 1	**Предварительный этап обработки набора данных**
Задачи	1. Загрузка данных в специализированное ПО. 2. Ознакомление с полученными данными. 3. Определение перечня возможных потенциальных проблем. 4. Формирование рабочих гипотез относительно данных. 5. Изучение специфики этапов обработки данных (приложение А. таблица 6)
Блок 2	**Основной этап обработки набора данных**
Задачи	1. Изучение свойств столбцов: Выделите столбец или столбцы, которые вас интересуют, и примените различные функции и инструменты Excel для получения информации о данных в этих столбцах. Например: • Вычислите среднее, медиану, минимальное и максимальное значения столбца с помощью функций, таких как AVERAGE, MEDIAN, MIN и MAX. • Рассчитайте общее количество значений или уникальных значений с помощью функций, таких как COUNT и COUNTIF. • Создайте гистограмму или диаграмму рассеяния для визуализации распределения данных.

Блок 2	Основной этап обработки набора данных
Задачи	2. Обнаружение пропущенных значений: Используйте функции Excel, такие как ISBLANK или COUNTBLANK, чтобы определить, есть ли пропущенные значения в данных. Примените фильтры или условное форматирование, чтобы выделить ячейки с пропущенными значениями для дальнейшего анализа. 3. Анализ выбросов и ошибочных данных: Примените функции Excel, такие как IQR (межквартильный размах) или Z-оценка, чтобы выявить выбросы или потенциально ошибочные значения в данных. Используйте условное форматирование, чтобы выделить ячейки с выбросами для дальнейшего рассмотрения. 4. Кодирование категориальных признаков: В Excel кодирование категориальных признаков может быть выполнено с использованием различных подходов. Вот некоторые из них: • Метод "Один против всех" (One-Hot Encoding): Создайте новый столбец для каждой уникальной категории в исходном признаке. Запишите в новые столбцы значения 0 или 1 в зависимости от принадлежности исходного признака к соответствующей категории. • Метод "Целочисленное кодирование" (Integer Encoding): Присвойте каждой уникальной категории в исходном признаке уникальное числовое значение. Замените значения в исходном столбце соответствующими числовыми значениями. • Метод "Замена на числовой код" (Numeric Code Replacement):

Блок 2	Основной этап обработки набора данных
Задачи	Создайте новый столбец, в котором каждой уникальной категории будет присвоено числовое значение. Замените значения в исходном столбце на соответствующие числовые коды. 5. Создание новых признаков: В Excel вы можете создавать новые признаки путем применения различных функций и операций над существующими данными. Вот некоторые способы создания новых признаков в Excel • Арифметические операции: Используйте арифметические операции, такие как сложение, вычитание, умножение и деление, для создания новых признаков на основе существующих. • Преобразование текстовых данных: Используйте функции текстового преобразования, такие как LEFT, RIGHT, MID, для извлечения части текстовой строки и создания нового признака на основе этой информации. Например, можно извлечь первый символ из текстового столбца и использовать его в качестве нового признака. • Функции Excel: Например, функция CONCATENATE позволяет объединять значения из нескольких столбцов в один новый столбец, функция IF позволяет создавать условные выражения для присвоения значений новому признаку на основе условий, а функция LEN позволяет определить длину значения в ячейке, что может быть полезно для создания признака длины.

Блок 3	Заключительный этап обработки набора данных
Задачи	1. Создание сводных таблиц: Используйте инструмент сводных таблиц в Excel для создания сводных таблиц, суммирующих и сгруппированных данных. Это позволяет получить общую сводку данных и провести более детальный анализ. 2. Визуализация данных: Используйте диаграммы и графики Excel для визуализации данных и выявления паттернов или трендов. Например, постройте гистограмму, круговую диаграмму или диаграмму рассеяния для визуализации распределения, соотношений или зависимостей между переменными. 3. Документирование результатов: Запишите свои наблюдения и выводы в отчёте.

Данный этап необходимо оформить в соответствии с примером, указанным в приложении Б.

Оценка качества набора данных

1. Оценка качества набора пространственных данных

Оценка качества набора пространственных данных основана на ГОСТ Р 57773-2017 «Пространственные данные. Качество данных», но не следует всем описанным положениям.

В данных методических рекомендация оценка качества используется для удостоверения в том, что набор данных отвечает уровням соответствия качества, установленным требованиями пользователя.

Первым этапом необходимо указать единицы качества данных.

Ниже приведены примеры того, что определяет область определения качества данных:

а) комплект наборов данных;

б) набор данных;

в) поднабор данных, определяемый одной или несколькими из следующих характеристик:

1) типы элементов (наборы типов объектов, атрибутов объектов, операций с объектами или отношений объектов);

2) конкретные элементы (наборы экземпляров объектов, значения атрибутов или экземпляры отношений объектов);

3) географическая протяженность;

4) временная протяженность (заданные временные рамки и точность временных рамок).

Рекомендуется в качестве единицы выбрать «набор данных».

Представить изображение набора данных, в виде векторного слоя наложенного на растровую подложку. Растровая подложка определяет «реальный мир», который обычно содержит больше объектов, чем содержится в наборе данных.

Во *втором этапе* необходимо выбрать меры качества. Перед выбором мер следует ознакомиться с основными элементами качества:

1) Полнота

Полнота определяется наличием и отсутствием объектов, их атрибутов и отношений. Она состоит из двух элементов качества данных:

Присутствие: избыточность данных в наборе данных;

Пример – Только здания с площадью больше 5 м должны быть включены в набор данных. Информация о наличии зданий до 5 м представляется в виде избыточности.

Отсутствие: отсутствие данных в наборе данных.

Пример – Отсутствие жилой недвижимости Англии или Уэльса в наборе данных.

2) Логическая согласованность

Под логической согласованностью понимают степень соответствия логических правил структуры данных, атрибутов и отношений (структура данных может быть концептуальной, логической или физической).

Логическая согласованность состоит из четырех элементов качества данных:

- концептуальная согласованность: соответствие правилам концептуальной схемы;
- доменная согласованность: соответствие значений атрибутов области допустимых значений;
- согласованность по формату: степень, с которой данные хранятся в соответствии с физической структурой набора данных;
- топологическая согласованность: корректность представления закодированных топологических характеристик набора данных.

3) Позиционная точность

Под позиционной точностью понимают точность положения объектов внутри пространственной системы координат.

Она состоит из трех элементов качества:

- абсолютная или внешняя точность: степень соответствия заявленных значений координат значениям координат, принятым в качестве правильных или являющимся правильными;
- относительная или внутренняя точность: степень соответствия относительного положения объектов в наборе данных их соответствующим исходным положениям, принятым в качестве правильных или являющимся правильными;
- позиционная точность матричных данных: соответствие значений пространственного позиционирования матричных данных значениям, принятым в качестве правильных или являющимся правильными.

4) Тематическая точность

Под тематической точностью понимают точность количественных атрибутов и корректность неколичественных атрибутов, классификаций объектов и их отношений.

Она состоит из трех элементов качества:

- правильность классификации: соответствие классов объектов или их атрибутов предметной области (например, реальной ситуации или эталонному набору данных);
- правильность неколичественных атрибутов: определение, является ли неколичественный атрибут правильным или неправильным;
- точность количественных атрибутов: степень соответствия значения количественного атрибута значению, принятому в качестве правильного или являющемуся правильным.

5) Временное качество

Под временным качеством понимают качество временных атрибутов и временных отношений объектов.

Оно состоит из трех элементов качества:

- точность измерения времени: степень соответствия заявленных временных измерений значениям, принятым в качестве правильных или являющимся правильными;
- согласованность по времени: правильность временного порядка событий.
- временная достоверность: достоверность данных по отношению ко времени.

Пример – 33 марта - пример неверных данных.

В соответствии с предложенными элементами качества необходимо выбрать по одной мере на каждый элемент. Перечень стандартизированных мер качества данных приведен в ГОСТ Р 57773-2017 «Пространственные данные. Качество данных» в приложении D.

Если какой-то элемент не рассматривается, то необходимо распределить меру для тех элементов, которые выбраны в соответствии со спецификой набора данных, но общее число мер должно равняться пяти.

Третий этап – определение метода оценки качества данных.

В стандарте представлено два метода оценки качества набора данных – прямой и косвенный. Рекомендуется использовать прямой метод оценки качества данных. Прямой метод делится на внутреннюю и внешнюю оценку. Внутренняя прямая оценка качества данных использует только те данные, которые содержатся в оцениваемом наборе данных. Внешняя прямая оценка качества требует применения эталонных данных, внешних по отношению к тестируемому набору данных.

Как для внешних, так и для внутренних методов оценки требуется произвести полный контроль. При полном контроле тестируется каждый элемент в генеральной совокупности, определенной областью качества данных.

В случае для внешнего метода оценки необходимо указать название эталонного набора и ссылку на набор (при наличии). Также требуется представить изображение этого набора на подобии с изображением указанным в первом этапе.

Четвертый этап – определение оценки качества данных на выходе.

В соответствии с выбранными мерами качества необходимо определить насколько они соответствуют требованиям.

Если показатель не прошел оценку, следует представить описание ошибки и способы повышения качества данных.

Пятый этап – обобщение результатов оценки качества.

Обобщенный показатель качества данных (ADQR) включает результаты оценки качества данных на основе различных элементов качества данных или различных областей определения качества данных.

Далее предложены два способа обобщения результатов оценки качества.

Однозначная оценка пригодности/непригодности

Каждому показателю качества данных, вовлеченному в вычисления, придается логическое значение, равное единице (1), если значение показателя соответствует требованиям, и нулю (0), если нет.

Обобщенный показатель качества определяется уравнением:

$$ADQR = v_1 \times v_2 \times v_3 \times \ldots \times v_n,$$

где n - число групп определения качества данных.

Если $ADQR$=1, то общее качество набора данных считается полностью соответствующим требованиям, а значит, пригодно. Если $ADQR$=0, то качество считается не соответствующим требованиям, а значит, непригодно.

Взвешенная оценка пригодности/непригодности

Каждому показателю качества данных, вовлеченному в вычисление, придается логическое значение, равное единице (1), если значение показателя соответствует требованиям, и нулю (0), если нет.

Кроме того, на основании значимости показателя для оценки качества в целом каждому из них присваивается весовое значение в интервале от 0 и 1 включительно.

Сумма всех весов должна равняться 1. Выбор весов является субъективным решением, принимаемым разработчиком данных или пользователем данных.

Обобщенное качество определяется уравнением:

$$ADQR = v_1 \times w_1 + v_2 \times w_2 + v_3 \times w_3 + \cdots + v_n \times w_n,$$

где n - число групп определения качества данных.

Из предложенных способов необходимо выбрать один и в выводе описать полученный результат.

2. Оценка качества набора данных без пространственной специфики

Оценка качества набора данных основана на ГОСТ Р 57773-2017 «Пространственные данные. Качество данных» и «DAMA-DMBOK: Свод знаний по управлению данными. Второе издание.

В таблице 4 представлены основные показатели, которые необходимо оценить.

Таблица 4. Показатели оценки качества набора данных

Измерение	**Замеры**	**Требования к качеству**
Полнота данных	Количество избыточных элементов	Количество избыточных элементов должно равняться 0
	Количество повторяющихся экземпляров объекта	Количество повторяющихся экземпляров объекта должно равняться 0
Согласованность	Согласованность с концептуальной схемой	В наборе данных могут присутствовать только типы объектов и атрибуты, определенные в Приложении Б, таб. 1.1
	Согласованность по формату	В соответствии со спецификацией формата
Логическое соответствие	Количество противоречивых значений	Количество противоречивых значений должно равняться 0
	Количество атрибутов, не имеющих расшифровки	Количество атрибутов, не имеющих расшифровки должно равняться 0

Дополнительно требуется представить еще один показатель, определяемый пользователем (практикантом).

Обобщенный результат оценки качества набора данных необходимо рассчитать в соответствии с методом, представленным в предыдущем подразделе в параграфе «Взвешенная оценка пригодности/непригодности». В соответствии с данным методом следует представить таблицу распределения весов. В выводе описать полученный результат.

Пример оформления раздела представлен в Приложении Б.

Подготовка набора данных к публикации

Подготовка набора данных к публикации делится на несколько этапов:

- Определение соответствия набора данных требованиям;
- Определение метаинформации набора данных;
- Определение условий пользования набором данных.

1. Требования к публикации набора открытых данных

Набор данных должен соответствовать требованиям, представленным в таблицах 5-6.

Таблица 5. Машиночитаемое представление открытых данных

а.	Открытые данные должны публиковаться в форматах CSV, XML, JSON: I. В формате CSV рекомендуется публиковать данные, имеющие плоскую табличную форму, при этом в содержании записи не допускается использование символа перевода строки. II. Сложные иерархические данные рекомендуется публиковать в форматах XML, JSON.
б.	В случае если размер набора имеет значительный объем данных (более 30 Мбайт), рекомендуется архивировать его с помощью алгоритма архивирования, имеющего спецификацию в виде открытого стандарта
в.	Для представления набора открытых данных, содержащих сведения из различных предметных областей, должны использоваться существующие форматы разметки типовых данных (schema.org, YMapsML, XAL и т.п.), имеющие опубликованную спецификацию
г.	Атрибуты каждого набора открытых данных должны иметь краткое англоязычное представление (в виде англоязычных имен или краткого текста транслитерации)
д.	Не допускается представление данных в неструктурированной форме, затрудняющей автоматическую обработку (например, недопустимо представление набора данных в виде бинарных данных, включенных в файл офисного документа)

Таблица 6. Требования к публикации открытых данных в машиночитаемых форматах CSV, XML, JSON

1)	**Требования к расположению файла набора открытых данных**
а.	Форматом файла набора открытых данных является CSV/XML/JSON
б.	Файл имеет название «data-<Версия набора>-structure-<Версия структуры>.csv/xml/json», где <Версия набора> это версия набора открытых данных в формате «ISO 8601», с точностью не ниже чем «День», и <Версия структуры> это версия соответствующей структуры набора открытых данных в формате «ISO 8601», с точностью не ниже чем «День»
2)	**Требования к формату файла набора открытых данных**

а.	• Для файла в формате CSV: соответствие стандарту RFC – «Common Format and MIME Type for Comma-Separated Values (CSV) Files»
	  ○ Дополнительные требования к CSV:
	  ▪ Разделителем полей является знак «,» (запятая);
	  ▪ Ограничителем строк является знак «"» (универсальная двойная кавычка);
	  ▪ Разделителем целой и дробной части чисел является знак «.» (точка).
	• Для файла в формате XML: соответствие стандарту W3C – «Extensible Markup Language (XML) 1.1 (Second Edition)»
	• Для файла в формате JSON: соответствие стандарту представленным на сайте – http://json.org/json-ru.html

б.	Кодировка файла – «UTF-8»

2. *Требования к представлению метаинформации набора открытых данных*

Метаинформация набора открытых данных состоит из следующих частей:

- паспорт набора открытых данных;
- структура набора открытых данных;
- другая информация, описывающая набор открытых данных.

2.1. Паспорт набора открытых данных

В первом этапе данного подраздела необходимо представить человекочитаемый (таб. 7) и машиночитаемый (таб. 8-9) форматы представления паспорта набора открытых данных.

Паспорт открытых данных представляет собой совокупность сведений о наборе открытых данных, необходимых для установления факта принадлежности набора открытых данных к той или иной тематической рубрике, его потенциальной пригодности для решения задач потребителя, а также установления адреса размещения, способа загрузки и последующей автоматической обработки набора открытых данных.

Паспорт набора открытых данных должен иметь четко заданную структуру следующего вида:

Таблица 7. Содержание паспорта набора открытых данных

1	Идентификационный номер
2	Наименование набора открытых данных
3	Описание набора открытых данных
4	Владелец набора открытых данных
5	Ответственное лицо
6	Телефон ответственного лица
7	Адрес электронной почты ответственного лица
8	Гиперссылка (URL) на открытые данные
9	Формат набора открытых данных
10	Описание структуры набора открытых данных
11	Дата первой публикации набора открытых данных
12	Дата последнего внесения изменений
13	Содержание последнего изменения
14	Дата актуальности набора данных
15	Ключевые слова, соответствующие содержанию набора данных
16	Гиперссылки (URL) на версии открытых данных
17	Гиперссылки (URL) на версии структуры набора данных
18	Версия методических рекомендаций

1) Идентификационный номер (код) набора открытых данных формируется следующим образом:

a) формат идентификационного номера: <код организации>-<наименование набора>;

b) код организации представляет собой идентификационный номер налогоплательщика (ИНН);

c) наименование набора открытых данных – сокращенное англоязычное название набора открытых данных, указывается в одно слово (уникальное в пределах организации).

Пример: 7712345678-showrooms.

2) Описание набора открытых данных

Данное поле заполняется самостоятельно публикатором, дополнительных требований не предъявляется.

3) Владелец набора данных

Указывается информация о владельце-организации публикуемого набора открытых данных.

4) Ответственное лицо

Указывается фамилия, имя и отчество, а также должность ответственного лица в формате: «Фамилия Имя Отчество, должность».

5) Телефон ответственного лица

Международный формат номера мобильного телефона: +7 *** ***-**-**.

Международный формат номера стационарного телефона: + < код страны > < код города или сети > < номер телефона >

6) Формат набора открытых данных

Формат данных указывается заглавными буквами без каких-либо дополнительных знаков, например: CSV.

В таблице 8 представлено описание требований к публикации паспорта набора открытых данных в машиночитаемых форматах CSV, XML, JSON.

7) Дата первой публикации набора открытых данных

Формат даты: ДД.ММ.ГГГГ.

8) Дата последнего внесения изменений

Формат даты: ДД.ММ.ГГГГ.

9) Содержание последнего изменения

Данное поле может принимать одно из следующих значений:

- Обновление паспорта: «номер поля»;
- Обновление файла открытых данных: описание обновления;
- Обновление файла структуры: описание обновления;
- Произвольный текст – произвольное текстовое описание в иных случаях.

10) Дата актуальности набора данных

Формат даты: ДД.ММ.ГГГГ, либо «По запросу».

Таблица 8. Требования к публикации паспорта набора открытых данных в машиночитаемых форматах CSV, XML, JSON

1)	**Требования к расположению файла паспорта набора открытых данных**
а.	Машиночитаемый формат паспорта набора открытых данных представлен отдельным файлом в формате CSV/XML/JSON
б.	Файл имеет название «meta.csv/xml/json»
2)	**Требования к формату файла паспорта набора открытых данных**

а.	• Для файла в формате CSV: соответствие стандарту RFC – «Common Format and MIME Type for Comma-Separated Values (CSV) Files»
	  ○ Дополнительные требования к CSV:
	  ▪ Разделителем полей является знак «,» (запятая);
	  ▪ Ограничителем строк является знак «"» (универсальная двойная кавычка);
	  ▪ Разделителем целой и дробной части чисел является знак «.» (точка).
	• Для файла в формате XML: соответствие стандарту W3C – «Extensible Markup Language (XML) 1.1 (Second Edition)»
	• Для файла в формате JSON: соответствие стандарту представленным на сайте – http://json.org/json-ru.html

б.	Кодировка файла – «UTF-8»

В таблице 9 представлено описание структуры данных паспорта набора открытых данных в машиночитаемом формате, где по горизонтали перечислены атрибуты полей паспорта набора открытых данных, по вертикали перечислены поля паспорта набора открытых данных.

Таблица 9. Структура данных паспорта набора открытых данных в машиночитаемом формате

Поле паспорта	**Наименование поля**	**Значение поля**
Версия методических рекомендаций	standardversion	Ссылка на версию методических рекомендаций, которой соответствует публикация этого паспорта набора открытых данных
Идентификационный номер	identifier	Идентификационный номер набора открытых данных
Наименование набора данных	title	Наименование набора открытых данных
Описание набора данных	description	Подробное описание набора открытых данных
Владелец набора данных	creator	Владелец набора открытых данных, юридическое или физическое лицо, которое публикует свои данные
Дата первой публикации набора данных	created	Дата первичной публикации набора открытых данных в формате «ISO 8601», с точностью не ниже чем «День»
Дата последнего внесения изменений	modified	Дата последнего внесения изменения в набор, структуру или паспорт ОД в формате «ISO 8601», с точностью не ниже чем «День» и достаточной для отделения актуальной версии набора ОД от предыдущей версии
Ключевые слова, соответствующие содержанию набора данных	subject	Список ключевых слов соответствующих содержанию набора, разделенных между собой знаком «,» (запятая)
Наименование набора данных	format	Формат набора открытых данных в нижнем регистре. Например: csv, xml

Таблица 9. Продолжение

Поле паспорта	**Наименование поля**	**Значение поля**
Формат данных	provenance	Подробное описание набора открытых данных
Содержание последних изменений	valid	Описание внесенных изменений в последнюю версию набора открытых данных
Дата актуальности	valid	Дата, до которой будет актуальной последняя версия набора открытых данных в формате «ISO 8601», с точностью не ниже чем «День»
Ответственное лицо	publishername	ФИО лица ответственного за публикацию текущего набора открытых данных
Телефон ответственного лица	publisherphone	Телефон лица ответственного за публикацию текущего набора открытых данных, в следующем формате: «+<код страны><код региона><номер телефона>»
Адрес электронной почты ответственного лица	publishermbox	Электронная почта лица ответственного за публикацию текущего набора открытых данных
Файл набора открытых данных	data-<data-version>-structure-<structure-version>	Ссылка на файл набора открытых данных
Файл структуры набора открытых данных	structure-<structure-version>	Ссылка на файл структуры набора открытых данных

Требование к структуре данных в машиночитаемом формате XML/JSON: Соответствие файлу структуры паспорта набора открытых данных – «meta-schema.xsd/json».

2.2. Структура набора открытых данных

В таблицах 10-11 представлено описание структуры и требования к структуре набора открытых данных. Во втором этапе данного раздела необходимо представить структуру набора открытых данных.

Таблица 10. Требования к публикации структуры набора открытых данных в машиночитаемых форматах CSV, XSD, JSON

1)	**Требование к структуре файла набора открытых данных**
а.	Файл набора открытых данных соответствует файлу структуры набора открытых данных
2)	**Требования к расположению файла структуры набора открытых данных**
а.	Файл структуры набора открытых данных представлен отдельным файлом в формате CSV/XSD/JSON
б.	Файл имеет название «structure-<Версия структуры>. csv/xsd/json», где <Версия структуры> это версия соответствующей структуры набора открытых данных в формате «ISO 8601», с точностью не ниже чем «День»
3)	**Требования к формату файла структуры набора открытых данных**
а.	• Для файла в формате CSV: соответствие стандарту RFC – «Common Format and MIME Type for Comma-Separated Values (CSV) Files» • Для файла в формате XSD: соответствие стандарту XML Schema – http://www.w3.org/TR/xmlschema-0/, http://www.w3.org/TR/xmlschema-1/, http://www.w3.org/TR/xmlschema-2/ • Для файла в формате JSON: соответствие стандарту представленным на сайте – http://json.org/json-ru.html
б.	Кодировка файла – «UTF-8»

Таблица 11. Требование к структуре файла структуры набора открытых данных

Атрибут поля реестра	**Заголовок атрибута**	**Значение атрибута**
Наименование поля набора открытых данных	field name	Наименование поля набора открытых данных
Английское описание поля набора открытых данных	english description	Описание поля набора открытых данных на английском языке
Русское описание поля набора открытых данных	russian description	Описание поля набора открытых данных на русском языке
Тип данных поля набора открытых данных	type	Тип данных поля набора открытых данных на английском языке

Общие требования к представлению результатов по подразделам 1-2:

- Представить название набора данных в соответствии с требованиями, представленными в таблице 6;
- Указать формат и объем набора данных;
- Представить содержание набора данных в приложении и в основном тексте указать ссылку на приложение;
- В приложении в зависимости от формата требуется:
    - набор данных в формате CSV можно представить в виде графического представления (снимок экран табличного процессора, ограниченной области) или в текстовой форме.
    - набор данных в формате XML или JSON необходимо представить в приложении в виде исходного кода.
- Сформировать паспорт набора открытых данных в соответствии с таблицей 7;

- В приложении представить машиночитаемый формат паспорта набора открытых данных:
    - в формате CSV можно представить в виде графического представления (снимок экран табличного процессора, ограниченной области) или в текстовой форме.
    - в формате XML или JSON необходимо представить в приложении в виде исходного кода.
- При формировании паспорта набора открытых данных в формате XML или JSON необходимо в приложении представить структуру паспорта набора открытых данных (meta-schema.xsd/json);
- В приложении представить структуру набора открытых данных:
    - в формате CSV можно представить в виде графического представления (снимок экран табличного процессора, ограниченной области) или в текстовой форме.
    - в формате XSD или JSON необходимо представить в приложении в виде исходного кода.

3. Условия использования набора открытых данных

Публиковать набор открытых данных рекомендуется с использованием открытой лицензии.

Условия использования не должны требовать от пользователей заключения какого-либо договора.

Условия использования не должны ограничивать потребителей открытых данных в применении данных наборов в некоммерческих и коммерческих целях.

Рекомендуется использовать лицензии, представленные в приложении А таб. 6.

Среди предложенных лицензий необходимо выбрать ту, которая применима к сформированному набору данных.

Пример оформления раздела представлен в Приложении Б.

Заключение

Практика завершается составлением отчета по выполненным работам. Отчет должен быть оформлен в соответствии со стандартом оформления печатных работ в МИИГАиК. В содержании должны быть отражены результаты работ в соответствии с выбранным тематическим планом.

В последний день практики необходимо представить следующие материалы:

1. Отчет в печатной форме в формате DOC/DOCX, PDF;
2. Набор данных в формате CSV/XML/JSON;
3. Паспорт набора данных в формате CSV/XML/JSON;
4. Структура набора данных в формате CSV/XSD/JSON.
5. Результат предварительной обработки набора данных в формате XLSX / Исходный код реализации предварительной обработки набора данных в формате IPYNB или PY.

Дополнительно в зависимости от специфики работ могут быть представлены следующие материалы:

1. Структура паспорта набора данных в формате XSD/JSON;
2. Набор справочных данных в формате CSV/XML/JSON.
3. Наборы пространственных данных следует сопроводить векторным слоем в формате GeoJSON/GeoPackage/GML.
4. Исходный код реализации автоматизированного сбора данных.

Приложение А

Таблица 1. Список некоторых источников открытых данных

№ п/п	Наименование источника	Ссылка на источник
1	Портал открытых данных Российской Федерации	https://data.gov.ru/
2	Портал открытых данных Правительства Москвы	https://data.mos.ru/
3	Система классификаторов Санкт-Петербурга	https://classif.gov.spb.ru/
4	Портал открытых данных Астраханской области	https://data.astrobl.ru/
5	Портал открытых данных Вологодской области	https://data.gov35.ru/
6	Портал открытых данных Краснодарского края	http://opendata.krasnodar.ru/
7	Портал открытых данных Липецкой области	https://opendata48.ru/datasets/

Таблица 2. Тематики наборов открытых данных, представленные из «Хартии открытых данных»

Категория данных	Примеры массивов данных
Компании	Компании/реестр предприятий
Преступность и правосудие	Статистика преступности, безопасность
Наблюдение за планетой	Метеорологические данные/сведения о погоде, сельском хозяйстве, лесоводстве, рыбной ловле и охоте
Образование	Список школ; результативность работы школ, цифровые навыки
Энергетика и окружающая среда	Уровни загрязнения, энергопотребление
Финансовые и контрактные вопросы	Заключенные сделки, подписанные контракты, поданные заявки на участие в тендере, будущие тендеры, местный бюджет, национальный бюджет (планируемый и расходуемый)
Геопространственные данные	Топография, почтовые индексы, национальные карты, местные карты
Международное развитие	Предоставление помощи, продовольственная безопасность, добывающая промышленность, землепользование
Подотчетность правительственного аппарата и демократия	Контактная информация для связи с правительством, результаты выборов, нормативно-законодательные акты и уставы, заработные платы (ставки заработной платы), знаки признательности/подарки
Здравоохранение	Данные о назначаемых препаратах, данные о результатах
Наука и исследования	Данные о геномах, исследовательская и образовательная деятельность, результаты экспериментов
Статистика	Национальная статистика, перепись,

	инфраструктура, уровень благосостояния, профессиональные навыки
Социальная мобильность и благосостояние	Жилищное обеспечение, медицинское страхование и пособие по безработице
Транспорт и инфраструктура	Расписание общественного транспорта, точки доступа к широкополосным каналам

Таблица 3. Документы, регулирующие публикацию данных

№ п/п	Наименование документа
1	Федеральный закон от 27 июля 2006 г. №149-ФЗ «Об информации, информационных технологиях и о защите информации»
2	Федеральный закон от 9 февраля 2009 г. №8-ФЗ «Об обеспечении доступа к информации о деятельности государственных органов и органов местного самоуправления»
3	Федеральный закон от 7 июня 2013 г. №112-ФЗ «О внесении изменений в Федеральный закон «Об информации, информационных технологиях и о защите информации» и Федеральный закон «Об обеспечении доступа к информации о деятельности государственных органов и органов местного самоуправления»
4	Распоряжение Правительства Российской Федерации от 10 июля 2013 г. №1187-р «О перечнях информации о деятельности государственных органов, органов местного самоуправления, размещаемой в сети «Интернет» в форме открытых данных»
5	Федеральный закон от 27.07.2006 N 152-ФЗ (ред. от 21.07.2014) «О персональных данных» (с изм. и доп., вступ. в силу с 01.09.2015)
6	Различные нормативные акты, определяющие ограничения, касающиеся отдельных видов деятельности – например, таких как профессиональная тайна врачей (Федеральный закон от 21.11.2011 N 323-ФЗ «Об основах охраны здоровья граждан в Российской Федерации»)

Таблица 4. Список табличных процессоров

Открытое ПО	Проприетарное ПО
LibreOffice Calc	Microsoft Office Excel
OpenOffice Calc	МойОфис

Таблица 5. Список мобильных ГИС

Таблица 6. Список открытых лицензий

Название лицензии	Символьное обозначение	Графическое обозначение	Пояснение
Creative Commons Attribution	CC BY		Разрешено свободное использование произведения, при условии указания его автора.
Creative Commons Attribution-ShareAlike	CC BY-SA		Разрешено свободное использование произведения, при условии указания его автора. Также все производные произведения, должны распространяться под лицензией CC BY-SA.
Creative Commons CCZero	CC0		Протокол, представляющий из себя лицензию, максимально приближенную к досрочному переводу произведения в

			общественное достояние. Протокол даёт возможность автору отказываться от всех авторских и смежных прав на произведение, а также от всех связанных с ними претензий и исков по отношению к произведению насколько это возможно в рамках законодательства. Соответственно, пользователи получают возможность свободно использовать такое произведение, не указывая даже имя автора.
Open Data Commons Attribution License	ODC-BY		Эта лицензия разрешает совместное использование и редактирование данных для любых целей. В качестве условия лицензии требуется указание автора произведения (https://opendatacommons.org/licenses/by/1-0/)
Open Data Commons Public Domain Dedication and License	PDDL		Аналог Creative Commons CCZero (https://opendatacommons.org/licenses/pddl/1-0/)

Таблица 7. Этапы предварительной обработки данных

Этап предварительной обработки данных	**Описание**
Загрузка данных	Получение данных из источника и загрузка их в рабочую среду для дальнейшей обработки
Изучение данных	Анализ структуры данных, изучение содержимого, идентификация признаков и общая оценка качества данных.
Обработка пропущенных данных	Определение и обработка пропущенных значений, включая удаление, заполнение или интерполяцию.
Обработка выбросов и ошибочных данных	Идентификация и обработка выбросов и ошибочных данных, таких как удаление или коррекция ошибок.
Нормализация данных	Приведение данных к одному масштабу или диапазону значений для облегчения сравнения и анализа.
Кодирование категориальных признаков	Преобразование категориальных признаков в числовой формат для использования в алгоритмах машинного обучения.

Уменьшение размерности	Снижение размерности данных для сокращения количества признаков и упрощения анализа.
Создание новых признаков	Создание новых признаков на основе имеющихся данных или применение преобразований для улучшения модели.
Масштабирование данных	Преобразование данных для сокращения влияния выбросов и улучшения производительности моделей.

Приложение Б

Пример оформления раздела «Концептуальное проектирование набора данных»

1. Концептуальное проектирование набора данных

Название набора данных: _____

_____

Описание набора данных: _____

_____

_____

_____

_____

_____

Таблица 1.1 Атрибутивные характеристики набора данных

Наименование поля	**Описание поля**	**Тип поля**	**Уник.**	**Обязат.**	**Длина**	**Колич. знаков после запятой**

Тематика, к которой может быть отнесен набор данных: _____

_____

Категория доступа: *общедоступная информация / информация ограниченного доступа*.*

** - требуется сделать дополнительное пояснение, что материал будет отредактирован, чтобы в подготовленной к публикации версии не было информации ограниченного доступа.*

Порядок распространения информации: *информация, распространяемая свободно / информация, распространение которой ограничено, запрещено или специально маркируется.*

Гражданско-правовой режим: *неохраняемый объект / охраняемый объект.*

Актуальность формирования данного набора: _____

_____

_____

_____

_____

Пример оформления раздела «Определение потенциальной востребованности набора открытых данных»

2. Определение потенциальной востребованности набора открытых данных

1.1. Определение популярности тематики

– Прямой способ

Количество заполненных анкет/опрошенных респондентов: _____

Таблица 2.1 Распределения баллов определения популярности тематик

№ п/п	Наименование тематики	Баллы по каждой анкете/опросу	Сумма баллов
1	Компании	*1,1,3,4,2*	*11*
2	Преступность и правосудие	*3,5,6,2,3*	*19*
…	…	…	…
14	Транспорт и инфраструктура	*2,3,2,4,5*	*16*
		Итого:	*451*

Сумма баллов, полученных тематикой: *19*

Общая сумма баллов для всех тематик: *451*

$$К_{тематики} = \frac{19}{451} = 0{,}04$$

** Для данного подраздела необходимо провести анализ полученного результата. Представить диаграммы, указывающие на характерный признак, повлиявший на результат (возраст респондентов, профессиональная деятельность и т.п.). Этот признак должен быть указан в заголовке диаграммы (Пример: «Большее количество респондентов, выставивших максимальный балл, старше 50 лет»)*

– Косвенный способ

Название источника открытых данных: _____

_____

Ссылка на источник: _____

_____

Таблица 2.1 Сопоставление категорий источника открытых данных с тематиками из «Хартии открытых данных»

№ п/п	Тематики «Хартии открытых данных»	Категории источника открытых данных	Количество наборов открытых данных	Среднее значение скачанных наборов	Среднее значение просмотров наборов
1	Компании	*Экономика и финансы*	*36*	*12 000*	*50 000*
2	Преступность и правосудие	-	-	-	-
…	…	…	…		
14	Транспорт и инфраструктура	*Транспорт*	*50*	*15 000*	*25 000*
		Итого:	*200*	*300 000*	*250 000*

Порядковый номер тематики, i: *14*

Кол-во наборов открытых данных данной тематики, n_i: *50*

Общее кол-во наборов открытых данных, $\sum n$: *200*

$$k_{14,1} = \frac{n_{14}}{\sum n} = \frac{10}{200} = 0{,}25$$

Среднее число скачанных наборов данной тематики, m_i: *15 000*

Общее число средних значений скачанных наборов данных, $\sum m$: *300 000*

$$k_{14,2} = \frac{m_{14}}{\sum m} = \frac{15\,000}{300\,000} = 0{,}05$$

Среднее число просмотров наборов данной тематики, l_i: *25 000*

Общее число средних значений просмотров наборов данных, $\sum l$:

250 000

$$k_{14,3} = \frac{l_{14}}{\sum l} = \frac{25\,000}{250\,000} = 0{,}10$$

Коэффициент популярности тематики, в которой определен набор данных:

$$К_{\text{тематики}} = \frac{k_1 + k_2 + k_3}{3} = \frac{0{,}25 + 0{,}05 + 0{,}10}{3} = 0{,}13$$

1.2. Анализ поисковых запросов

Название поискового сервиса: *Яндекс.Wordstat*

Количество запросов по названию формируемого набора, v: *34 000*

Количество запросов по названию «открытые данные», w: *118 000*

$$К_{\text{востр1}} = \frac{v}{w} = \frac{34\,000}{118\,000} = 0{,}29$$

1.3. Востребованность набора данных по результатам опроса/анкетирования

Кол-во анкет/опрошенных респондентов: *15*

Таблица 2.2 Количество анкет по выставленным баллам набору данных

Баллы:	1	2	3	4	5
Кол-во анкет:	*0*	*2*	*4*	*7*	*2*
Сумма баллов:	*0*	*4*	*12*	*28*	*10*

Сумма всех баллов, выставленных набору данных, p: *54*

Сумма баллов по каждой заполненной анкете за 4 и 5 баллы, t: *38*

Коэффициент востребованности набора данных по результатам опроса/анкетирования:

$$К_{\text{востр2}} = \frac{t}{p} = \frac{38}{54} = 0{,}70$$

1.4. Расчет общего коэффициента потенциальной востребованности набора данных

$$К_{востр\ общ} = \frac{К_{востр1} + К_{востр2} + К_{тематики}}{3} \times 100\%$$

$$= \frac{0{,}29 + 0{,}70 + 0{,}13}{3} \times 100\% = 37\%$$

Группа потенциальной востребованности набора данных: *В*

Предложения по повышению класса востребованности набора данных*:

_____

_____

_____

_____

_____

_____

_____

Пример оформления раздела «Создание набора данных с помощью программных средств»

3. Создание набора данных с помощью программных средств

– Программные средства записи данных

Название программного средства записи данных: _____

_____

Количество записей: _____

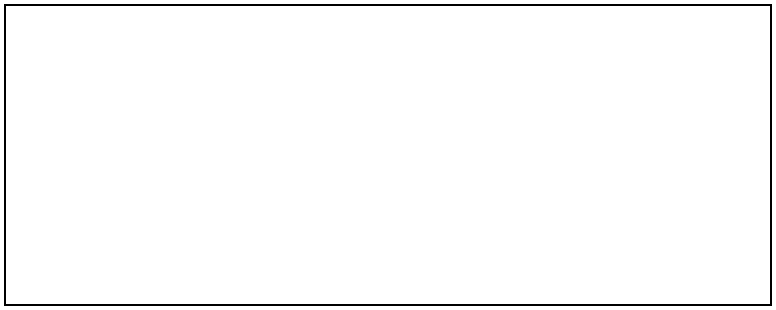


Рисунок 3.1 – Набор данных «*Название*», представленный в табличном формате

– Мобильные средства для сбора пространственных данных

Название района работ: _____

_____

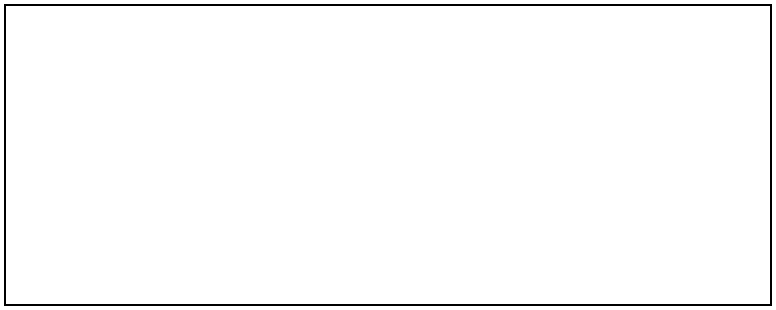


Рисунок 3.1 – Очерченный район работ

Количество объектов: _____

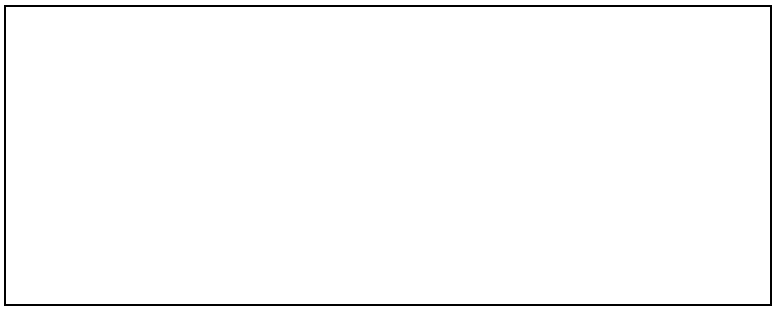


Рисунок 3.2 – Маршрут проведения работ

Название мобильной ГИС: _____

_____

Объяснение выбора данной ГИС:_____

_____

_____

_____

_____

Инструменты, функциональные возможности, использованные при создании набора данных:

1._____

2._____

3._____

4._____

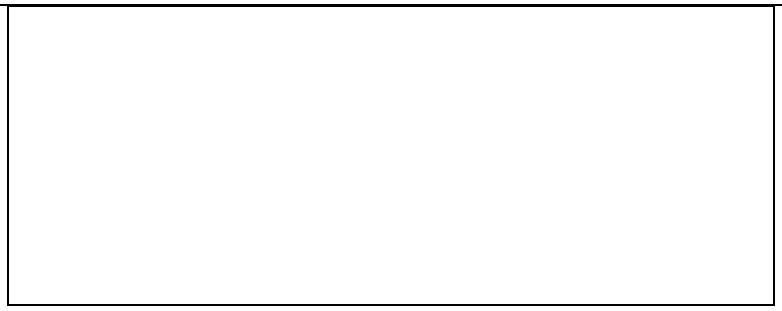


Рисунок 3.3 – Картографическая визуализация набора данных «*Название»* в мобильной ГИС «*Название»*

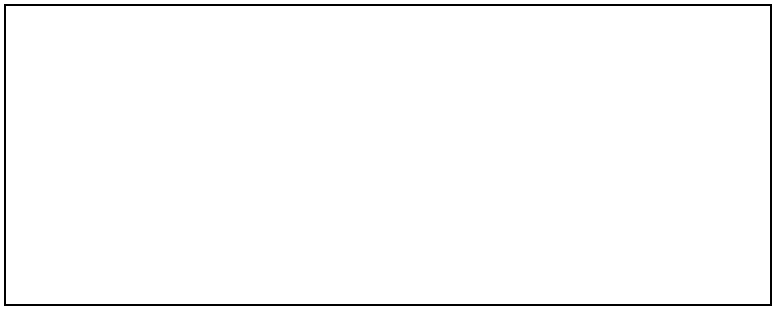


Рисунок 3.4 – Табличная визуализация набора данных «*Название*» в мобильной ГИС «*Название*»

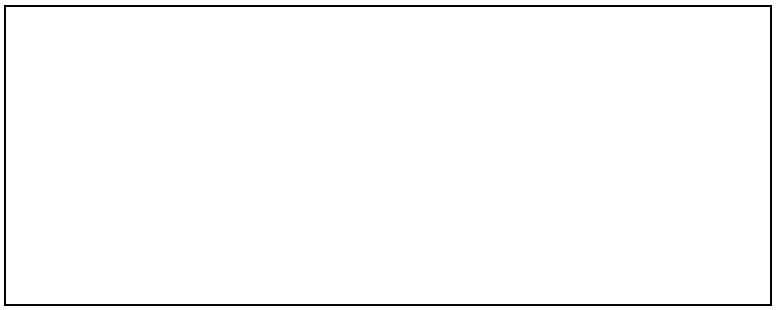


Рисунок 3.5 – Табличная визуализация набора данных «*Название*» в «*Название табличного процессора*»

Дополнительные сведения:_____

_____

_____

_____

_____

Примечание: в дополнительные сведения вносится информация из таблицы 2 пункта 3 подпунктов b и c

– Автоматизированный сбор данных

Название сайта: _____

Ссылка на сайт: _____

Вывод об архитектуре сайта:

1. Подход к реализации структуры страницы сайта:_____

_____

2. Указать теги содержащие, интересующие данные:_____

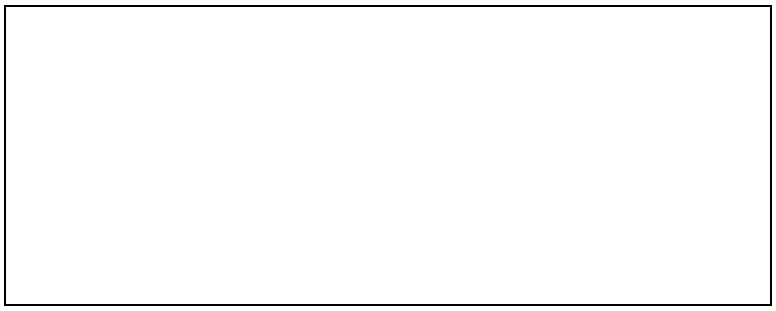
_____

3. Дополнительные сведения:_____

_____

Программные средства автоматизации:_____

_____



* Рисунок 3.1 – Графическое представление фрагмента данных неудовлетворяющих проверке

Исходный код реализации автоматизированного сбора данных представлен в приложении *Б*.

Пример оформления раздела «Предварительная обработка набора данных»

4. **Предварительная обработка набора данных**

– Программные средства обработки данных

Название программных средств обработки данных: _____

_____

– Ознакомление с данными

Таблица 4.1 Исследование данных

Название элемента	**Название характеристики**	**Диапазон значений**	**Графическое представление распределения данных**
Здание	*Этажность*	*Диапазон значений*	*Гистограмма распределения значений (для текстовых*

		данных	*значений распределение по категориям)*

Анализ характеристик:_____

_____

_____

_____

Перечень проблем с данными: _____

_____

_____

_____

Предложение по устранению аномалий: _____

_____

_____

_____

_____

– Реализация процесса очистки данных

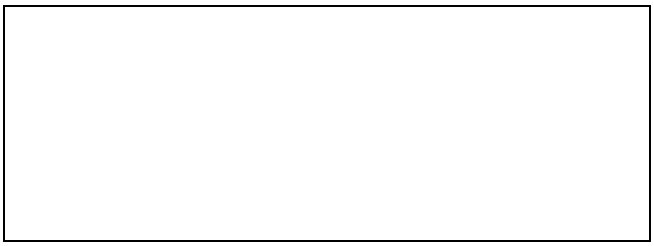


Рисунок 4.1 – Пример проблемы с характеристикой «*Название*», представленный в табличном формате/ гистограммы/диаграммы рассеивания

Описание проблемы:_____

_____

Название примененного инструмента:_____

_____

Выводы по результатам примененного инструмента:_____

_____

_____

_____

_____

_____

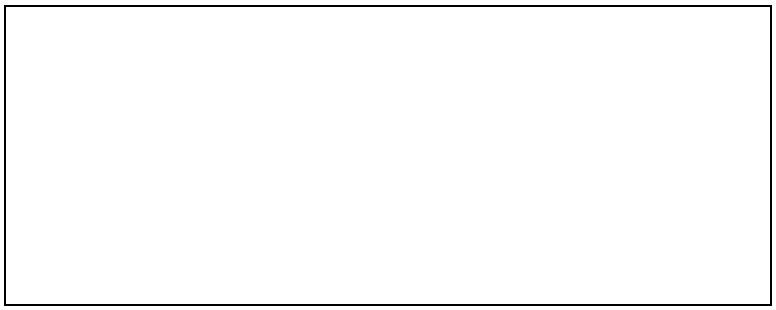


Рисунок 4.2 – Визуализация исправленной характеристики «*Название*», представленный в виде гистограммы/диаграммы рассеивания

– Обработка данных для машинного обучения

Таблица 4.2 Таблица Кодирование категориальных признаков

Название элемента	**Название характеристики**	**Метод кодирования**	**Исходные значения**	**Целевые значения**
Здание	*Тип строения*	*Название метода*	*Значения в исходном наборе данных*	*Новое значение*

Дополнительные сведения: _____

_____

_____

_____

Таблица 4.3 Создание синтетических признаков

Название элемента	**Название исходных характеристик**	**Название синтетического признака**	**Формула расчета**
Здание	*Этажность*	*Высота здания*	*Этажность * коэффициент = высота*

Возможное применение синтетического признака: _____

_____

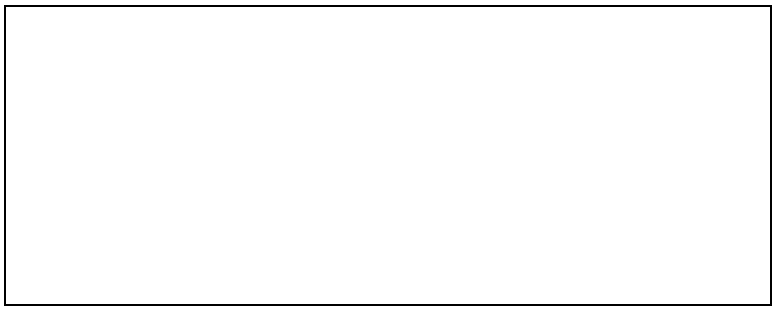


Рисунок 4.3 – Итоговый набор данных «*Название*», представленный в табличном формате

Пример оформления раздела «Оценка качества набора данных»

5. Оценка качества набора данных

– Оценка качества набора пространственных данных

Единица качества данных: *набор данных*

* Требования к качеству (при наличии спецификации):

Деревья:

1. *Максимально пропущено 10%.*
2. *Максимальный процент деревьев, которые могут иметь неправильную высоту, - 20%.*

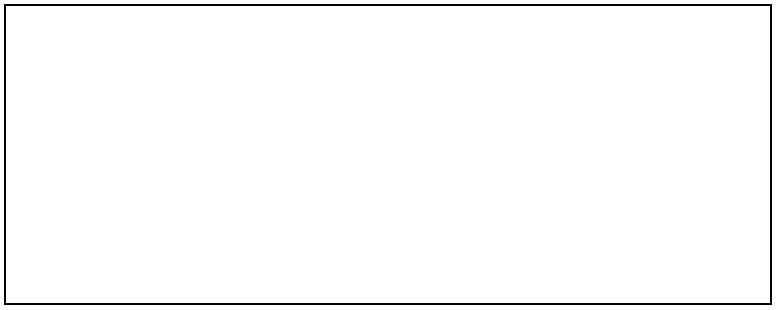


Рисунок 5.1 – Графическое представление набора данных**

** - *в графическом представлении необходимо указать объекты с подписями и с условным обозначением*

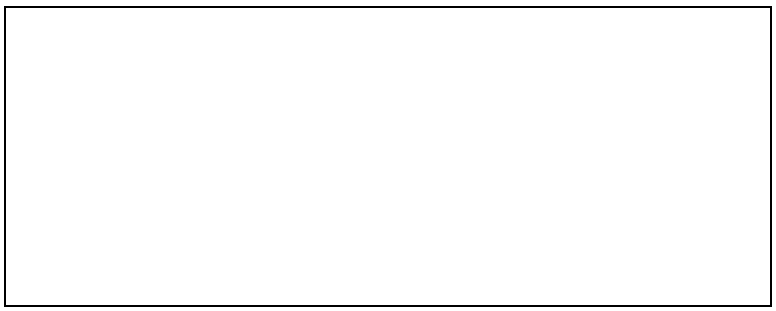
Таблица 5.1 Определение мер качества

Название элемента качества данных	Название меры качества, №
Логическая согласованность	Несогласованность с концептуальной схемой, 8
Полнота	Количество избыточных элементов, 2
Позиционная точность	Среднее значение позиционной неопределенности, 28
Временное качество	Количество несогласованных объектов со своей областью значений, 16
Тематическая точность	Количество неправильно классифицированных объектов, 60

Процедура оценки качества данных: *прямая внутренняя оценка; полная проверка*

* Название эталонного набора данных (при прямой внешней оценки):__ _____

* Ссылка на эталонный набор данных (при наличии):_____ _____



* Рисунок 5.2 – Графическое представление эталонного набора данных**

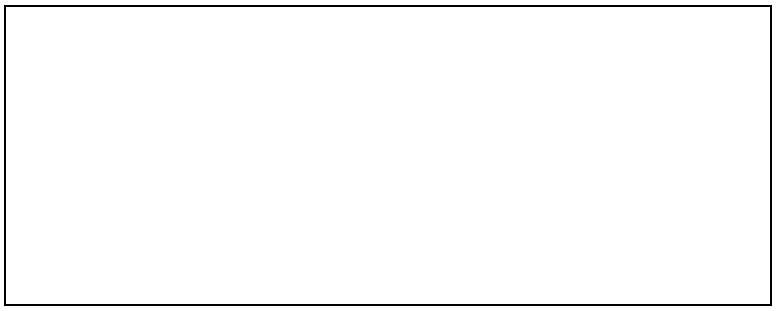
** - *в графическом представлении необходимо указать объекты с подписями и с условным обозначением*

Показатель № 1 *Количество избыточных элементов*

Таблица 5.2 Оценка *количества избыточных элементов*

Название элемента	**Количество элементов в наборе данных**	**Количество элементов в предметной области (эталонном наборе)**	**Количество избыточных элементов**	**Требования к качеству данных**	**Оценка, v_1**
Здание	*10*	*10*	*0*	*кол-во избыточных элементов должно*	*1*

				равняться 0	



* Рисунок 5.3 – Графическое представление ошибки по показателю № 1

* Описание ошибки:_____

_____

* Предложения по повышению качества:_____

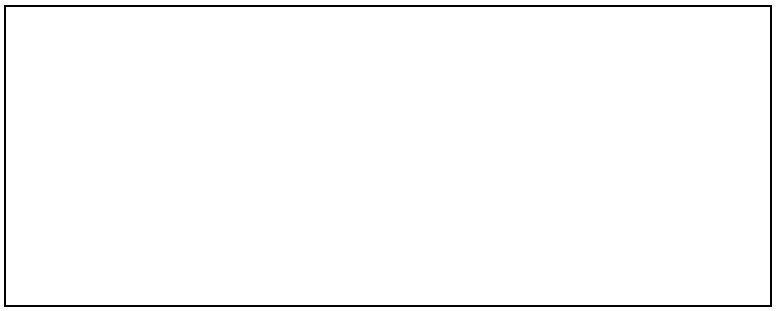
_____

_____

Показатель № 2 *Концептуальная согласованность*

Таблица 5.2 Оценка *несогласованности с концептуальной схемой*

Область определения	**Требования к качеству данных**	**Значение**	**Оценка,** v_2
Набор данных	*в наборе данных могут присутствовать только типы объектов и атрибуты, определенные в таб. 1.1*	*False (0) – соответствует концептуальной схеме*	*1*



* Рисунок 5.4 – Графическое представление ошибки по показателю № 2

* Описание ошибки:_____

_____

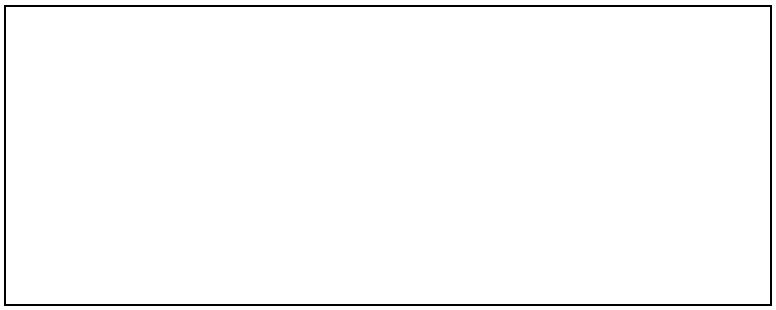
* Предложения по повышению качества:_____

_____

Показатель № 3 *Среднее значение позиционной неопределенности*

Таблица 5.3 Оценка *среднего значения позиционной неопределенности*

Средняя позиционная неопределенность горизонтальной абсолютной позиции, $\bar{e}$	**Требования к качеству данных**	**Оценка, v_3**
2,356 м	$\bar{e} < 5$ м	*1*



* Рисунок 5.5 – Графическое представление ошибки по показателю № 3

* Описание ошибки:_____

_____

* Предложения по повышению качества:_____

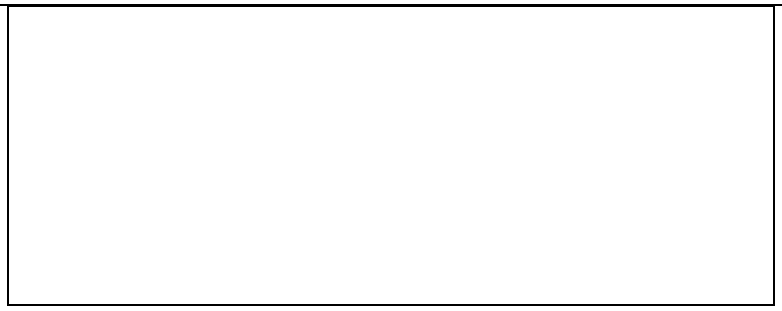
_____

_____

Показатель № 4 *Временная достоверность*

Таблица 5.4 Оценка *количества несогласованных объектов со своей областью значений*

Количество несогласованных объектов с форматом записи даты	**Требования к качеству данных**	**Оценка, v_4**
0	*количество несогласованных объектов с форматом записи даты должно равняться 0*	*1*



* Рисунок 5.6 – Графическое представление ошибки по показателю № 4

* Описание ошибки:_____

_____

* Предложения по повышению качества:_____

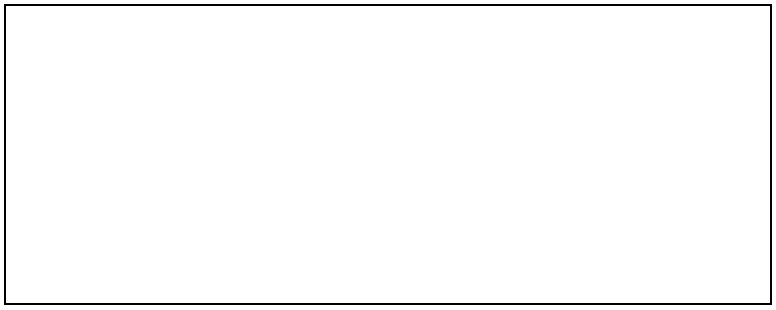
_____

_____

Показатель № 5 *Количество неправильно классифицированных объектов*

Таблица 5.5 Оценка *количества неправильно классифицированных объектов*

Количество неправильно классифицированных объектов	**Требования к качеству данных**	**Оценка, v_5**
1	*допустимое значение неправильно классифицированных объектов не больше 3*	*1*



* Рисунок 5.7 – Графическое представление ошибки по показателю № 5

* Описание ошибки:_____

_____

* Предложения по повышению качества:_____

_____

_____

-- *Однозначная оценка пригодности набора данных*:

$$ADQR = v_1 \times v_2 \times v_3 \times v_4 \times v_5 = 1 \times 1 \times 1 \times 1 \times 1 = 1$$

Вывод:_____

_____

_____

-- *Взвешенная оценка пригодности набора данных:*

Таблица 5.6 Распределение весов

Показатель	**Вес показателя**
v_1	*0,30*
v_2	*0,15*
v_3	*0,30*
v_4	*0,15*
v_5	*0,10*
Сумма весов:	*1,00*

$$ADQR = v_1 \times w_1 + v_2 \times w_2 + v_3 \times w_3 + v_4 \times w_4 + v_5 \times w_5$$
$$= 1 \times 0{,}30 + 1 \times 0{,}15 + 1 \times 0{,}30 + 1 \times 0{,}15 + 1 \times 0{,}10 = 1$$

Вывод:_____

_____

_____

– Оценка качества набора данных

* *Данный способ предназначен для набора данных без пространственной специфики*

1. Полнота данных

1.1. Количество избыточных элементов

Таблица 5.1 Оценка *количества избыточных элементов*

Количество элементов в наборе данных	**Количество записей в предметной области (эталонном наборе)**	**Количество избыточных элементов, *k***	**Требования к качеству данных**	**Оценка, v_1**
10	*10*	*0*	*k = 0*	*1*

* Описание ошибки:_____

_____

* Предложения по повышению качества:_____

_____

_____

1.2. Количество повторяющихся экземпляров объекта

Таблица 5.2 Оценка *количества повторяющихся экземпляров объекта*

Количество повторяющихся экземпляров объекта, u	**Требования к качеству данных**	**Оценка, v_2**
0	*$u = 0$*	*1*

* Описание ошибки:_____

_____

* Предложения по повышению качества:_____

_____

_____

2. Согласованность

2.1. Согласованность с концептуальной схемой

Таблица 5.3 Оценка *концептуальной согласованности*

Требования к качеству данных	**Значение**	**Оценка, v_3**
в наборе данных могут присутствовать только типы объектов и атрибуты, определенные в таб. 1.1	*True (1) – соответствует концептуальной схеме*	*1*

* Описание ошибки:_____

_____

* Предложения по повышению качества:_____

_____

_____

2.2. Согласованность по формату

Таблица 5.4 Оценка *конфликтов физической структуры*

Требования к качеству данных	**Значение**	**Оценка, v_4**
в соответствии со спецификацией формата CSV - RFC 4180	*False (0) – отсутствие конфликта физической структуры*	*1*

* Описание ошибки:_____

_____

* Предложения по повышению качества:_____

_____

_____

3. Логическое соответствие

3.1. Количество противоречивых значений

Таблица 5.5 Оценка *количества противоречивых значений*

Количество противоречивых значений, q	**Требования к качеству данных**	**Оценка, v_5**
0	$q = 0$	*1*

* Описание ошибки:_____

_____

* Предложения по повышению качества:_____

_____

_____

3.2. Контекстуальная ясность

Таблица 5.6 Оценка *количества атрибутов, не имеющих расшифровки*

Количество атрибутов, не имеющих расшифровки, o	**Требования к качеству данных**	**Оценка, v_6**
0	$o = 0$	*1*

* Описание ошибки:_____

_____

* Предложения по повышению качества:_____

_____

_____

4. *Показатель, определяемый пользователем*

…

5. Общий результат оценки набора данных

Таблица 5.8 Распределение весов

Показатель	**Вес показателя**
v_1	*0,25*
v_2	*0,15*
v_3	*0,25*
v_4	*0,10*
v_5	*0,10*
v_6	*0,10*
v_7	*0,05*
Сумма весов:	*1,00*

$$ADQR = v_1 \times w_1 + v_2 \times w_2 + v_3 \times w_3 + v_4 \times w_4 + v_5 \times w_5 + v_6 \times w_6 + v_7 \times w_7$$
$$= 1 \times 0{,}25 + 1 \times 0{,}15 + 1 \times 0{,}25 + 1 \times 0{,}10 + 1 \times 0{,}10 + 1 \times 0{,}10 + 1 \times 0{,}05 = 1$$

Вывод:_____

_____

_____

Пример оформления раздела «Подготовка набора данных к публикации»

6. Подготовка набора данных к публикации

Название набора данных: _____

_____

Формат набора данных: _____

Объем набора данных: _____

*Дополнительные сведения: *открытые данные имеют значительный размер, рекомендуется архивировать его с помощью алгоритма архивирования LZMA или LZMA2.*

Таблица 6.1 Паспорт набора открытых данных

1	Идентификационный номер	7701012399-*showrooms*
2	Наименование набора открытых данных	*Выставочные залы*
3	Описание набора открытых данных	*Сведения о выставочных залах с описанием деятельности, его расположения, контактных данных и времени его работы*
4	Владелец набора открытых данных	МИИГАиК
5	Ответственное лицо	*Иванов Иван Иванович*
6	Телефон ответственного лица	*+7 999 999-99-99*
7	Адрес электронной почты ответственного лица	*ivanov@example.com*
8	Гиперссылка (URL) на открытые данные	*data-2023-06-02-structure-2023-06-02.csv*
9	Формат набора открытых данных	*CSV*
10	Описание структуры набора открытых данных	*structure-2023-06-02.csv*
11	Дата первой публикации набора открытых данных	*02.06.2023*
12	Дата последнего внесения изменений	-
13	Содержание последнего изменения	-
14	Дата актуальности набора данных	*По запросу*
15	Ключевые слова, соответствующие содержанию набора данных	*Выставка, зал, выставочный зал, культура*
16	Гиперссылки (URL) на версии	*data-2023-06-02-structure-2023-06-*

	открытых данных	*02.csv (1)*
17	Гиперссылки (URL) на версии структуры набора данных	*structure-2023-06-02.csv (1)*
18	Версия методических рекомендаций	1.0

В приложении *А* представлены следующие сведения по данному разделу:

- текстовая форма набора данных в формате CSV;
- графическое представление паспорта набора данных в формате CSV;
- графическое представление структуры набора данных в формате CSV.

Выбор лицензии: *Creative Commons CCZero (CC0)*

Приложение В

Отчет о времени выполнения разделов/работ

Название раздела/работы	**Название подраздела**	**Время выполнения раздела/работы, ч**
Концептуальное проектирование набора данных		
Раскрытие набора данных		
Определение потенциальной востребованности набора открытых данных		
Создание набора данных с помощью программных средств	Программные средства записи данных	
	Мобильные средства для сбора пространственных данных	
	Автоматизированный сбор данных	
Предварительная обработка набора данных		
Оценка качества набора данных	Оценка качества набора пространственных данных	
	Оценка качества набора данных	
Подготовка набора данных к публикации		