|  |
| --- |
| [Введите название организации] |
| Декодер данных ЕМИСС |
| [Введите подзаголовок документа] |

|  |
| --- |
| Дарья Буланцева, Илья Гуленков, Дмитрий Фещенко, Валерия Власова  Москва, 2022 |

Оглавление

[Введение 2](#_Toc98708529)

[get\_structure(id, force\_upd=False) 3](#_Toc98708530)

[query\_size(filterdata) 4](#_Toc98708531)

[make\_query(filterdata) 4](#_Toc98708532)

[parse\_sdmx(response, id, nowrite=False) 4](#_Toc98708533)

[write\_db(id, order, colnames, fields\_id, fields\_title, fields\_values, fields\_codes, df) 5](#_Toc98708534)

[query\_splitter(filters, id, nowrite=False) 5](#_Toc98708535)

[load\_data(id) 6](#_Toc98708536)

[get\_data(id, force\_upd=False) 6](#_Toc98708537)

[get\_periods(id) 6](#_Toc98708538)

[monetary\_value() 7](#_Toc98708539)

[monthly\_introduction() 7](#_Toc98708540)

[monthly\_prices() 7](#_Toc98708541)

[user\_interface() 7](#_Toc98708542)

# Введение

Данная программа призвана облегчить выгрузку и распаковку открытых данных, доступных в Единой Межведомственной Информационной Системе (ЕМИСС) Fedstat.ru. Мотивация для создания этой утилиты состоит в следующем:

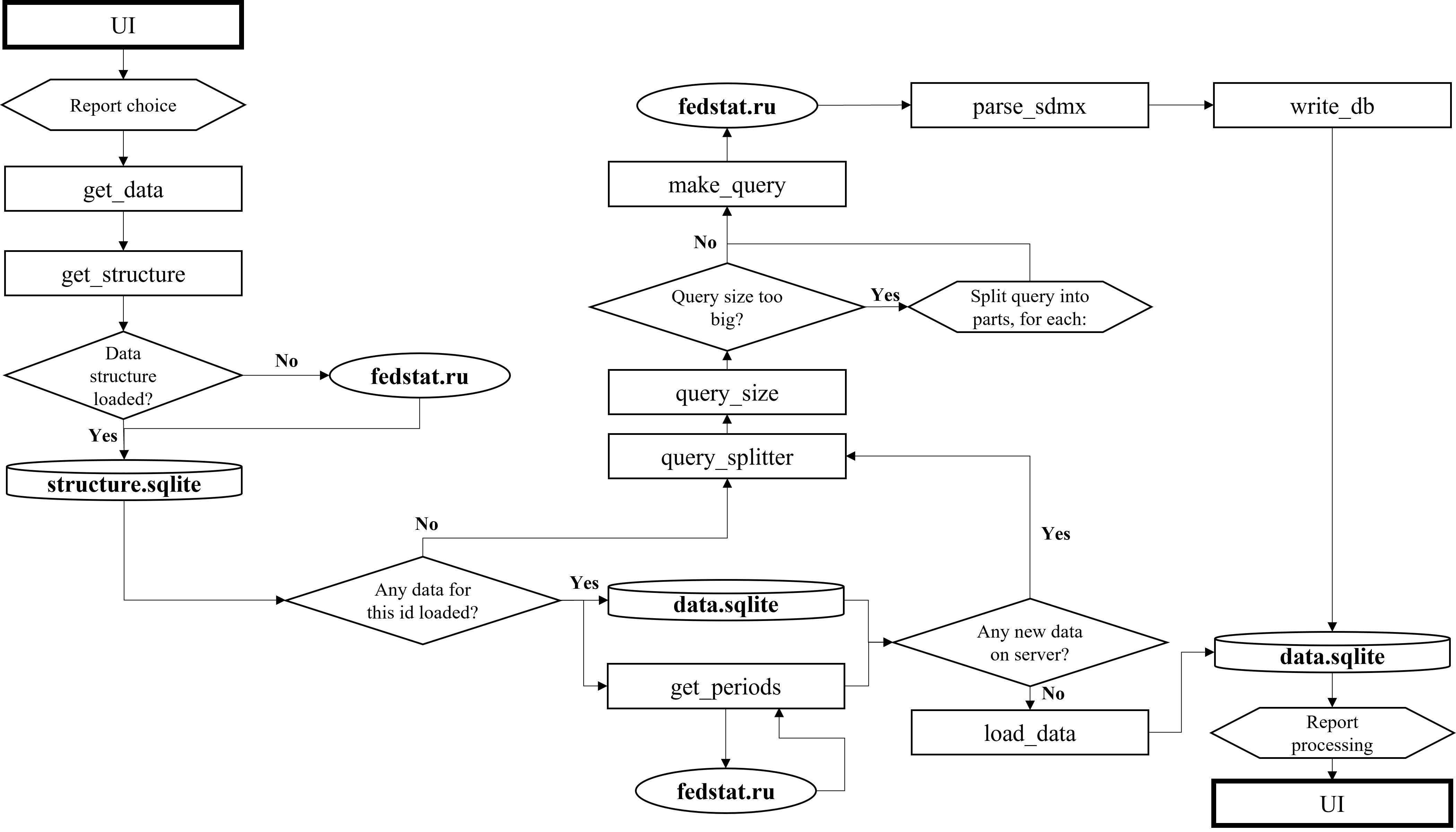
* Ресурс fedstat.ru демонстрирует крайне низкую устойчивость даже к умеренной нагрузке на сервер, что признается большей часть сообщества аналитиков, работающих с данными государственной статистики. В результате необходимые показатели зачастую оказываются недоступными для использования. Данная утилита позволяет решить эту проблему путем создания локальной копии данных на устройстве пользователя.
* Структура данных, отдаваемая системой fedstat.ru в пользовательский интерфейс «по умолчанию», крайне разнородна даже в показателях, характеризующих сходные социально-экономические процессы. В частности, один и тот же параметр (например, единица измерения) может быть задан как «фильтр» в одном показателе и как «группирующий параметр» в другом. Это значительно усложняет обработку и сведение информации из нескольких показателей между собой, поскольку требует предварительной унификации данных в ручном режиме (с использованием интерфейса fedstat или Excel).

Данная утилита позволяет решить описанные проблемы путем создания локальной копии данных, доступных по данному показателю. Для ее работы требуется знание уникального идентификатора набора данных, который можно получить из ссылки на него на сайте системы (fedstat.ru/indicator/34118). Вся обработка данных происходит автоматически, все доступные данные выгружаются в базу данных SQL, также реализована проверка актуальности локальной копии базы данных и, при необходимости, догрузки недостающих данных в нее.

Концептуально, утилита может работать с любыми[[1]](#footnote-1) данными из системы fedstat. Тем не менее, для целей отладки и генерации отчетов выбраны два набора данных о конъюнктуре рынка недвижимости регионов РФ (fedstat.ru/indicator/34118 и fedstat.ru/indicator/31452).

Пользователю предлагается выбрать данные для работы, после чего программа автоматически проверяет наличие и актуальность данных в локальной копии базы и скачивает/обновляет данные в базу данных SQL и файл формата csv. Данные представляются пользователю в удобном для понимания виде, коды регионов заменяются текстовыми расшифровками.

Ниже представлена стилизованная блок-схема, отражающая принцип работы утилиты и входящих в нее функций.



## get\_structure(id, force\_upd=False)

Обрабатывает структуру данных (фильтры). Структура уже скачанных данных выгружается из базы данных structure.sqlite или записывается в нее, если выбранная база данных ранее не скачивалась.

Аргументы:

**id** – строковое значение. Номер выгружаемой базы данных по кодировкам ЕМИСС.

**force\_upd** – значение типа bool. Соответствует требованию принудительно обновить данные в базе SQL. Значение по умолчанию False – не обновлять данные принудительно.

Возвращает:

**filters** –объект класса pandas.core.frame.DataFrame с данными о фильтрах, используемых Росстатом и впоследствии в базе данных.

Структура данных необходима для составления запросов на сайт ЕМИСС.

## query\_size(filterdata)

Определяет размер запроса, необходимый для выгрузки данных, на основе данных о числе значений фильтров в базе.

Аргументы:

**filterdata** – объект класса pandas.core.frame.DataFrame с данными о фильтрах, используемых ЕМИСС и впоследствии в базе данных.

Возвращает:

**S** –integer, размер запроса, количество значений фильтров.

Размер запроса необходим для определения количества запросов, необходимых для быстрой загрузки данных. Если значений фильтров слишком много, программа загружает данные несколькими запросами, число которых определяется функцией query\_splitter().

## make\_query(filterdata)

Переводит данные о фильтрах в формат, используемый сервером, принимающим запрос.

Аргументы:

**filterdata** – объект класса pandas.core.frame.DataFrame с данными о фильтрах, используемых ЕМИСС и впоследствии в базе данных.

Возвращает:

**query\_json** –словарь (dict) с данными о фильтрах, используемых в базе данных.

## parse\_sdmx(response, id, nowrite=False)

Обрабатывает файл данных, скачанный с сервера ЕМИСС. Обрабатывает повторяющиеся значения фильтров, записывает данные в базу данных через обращение к функции write\_db().

Аргументы:

**response** – объект типа urllib.response, содержащий данные, полученные с сервера ЕМИСС.

**nowrite** – значение типа bool. Соответствует требованию записать обработанные данные в базу SQL. Значение по умолчанию False – записать данные в базу.

Возвращает:

**parsed** –объект класса pandas.core.frame.DataFrame с обработанными данными.

Если nowrite=False, данные записываются в базу вызовом write\_db(), иначе возвращается датафрейм parsed.

## write\_db(id, order, colnames, fields\_id, fields\_title, fields\_values, fields\_codes, df)

Записывает обработанные данные в реляционную базу данных data.sqlite. Основные значения записываются в таблицу Data\_(индекс базы), для значений фильтров создаются дополнительные таблицы.

Аргументы:

**id** – строковое значение. Номер выгружаемой базы данных по кодировкам ЕМИСС.

**order** – объект типа list, содержащий названия столбцов в необходимом порядке.

**colnames** – объект типа list, содержащий названия будущих столбцов (столбцы из order + столбцы для time и value).

**fields\_id** – объект типа list, содержащий id фильтров.

**fields\_title** – объект типа list, содержащий названия фильтров.

**fields\_values** – объект типа list, содержащий значения фильтров.

**fields\_cоdes** – объект типа list, содержащий id значений фильтров.

**df** – объект типа list, элементами которого также являются списки. Содержит данные, которые необходимо записать в базу.

Результатом работы функции является запись данных в базу.

## query\_splitter(filters, id, nowrite=False)

Разделяет запрос на несколько в случае, если изначальный запрос слишком велик.

Аргументы:

**filters** – объект класса pandas.core.frame.DataFrame с данными о фильтрах, используемых ЕМИСС и впоследствии в базе данных.

**id** – строковое значение. Номер выгружаемой базы данных по кодировкам ЕМИСС.

**nowrite** – значение типа bool. Соответствует требованию записать обработанные данные в базу SQL. Значение по умолчанию False – записать данные в базу.

Возвращает:

**parsed** –объект класса pandas.core.frame.DataFrame с обработанными данными.

Возвращает датафрейм, когда nowrite=True, в остальных случаях опосредовано через функцию parse\_sdmx() данные записываются либо в базу, либо в датафрейм.

## load\_data(id)

Выгружает данные выбранного пользователем показателя из базы данных. Вызов этой функции происходит, когда в базе содержатся все последние данные, опубликованные ЕМИСС для этого показателя.

Аргументы:

**id** – строковое значение. Номер выгружаемой базы данных по кодировкам ЕМИСС

Возвращает:

**[result, titles]** – список, элементами которого являются датафрейм results, содержащий данные выбранного показателя, а также список titles с названиями фильтров, используемых для этого показателя.

Полученные данные затем используются для составления отчетов.

## get\_data(id, force\_upd=False)

Загружает данные выбранного пользователем показателя для составления отчетов. Во время вызова этой функции данные выбранного показателя опосредовано через другие функции либо загружаются с сервера ЕМИСС, когда они отсутствуют в базе, либо загружаются из самой базы.

Аргументы:

**id** – строковое значение. Номер выгружаемой базы данных по кодировкам ЕМИСС.

**force\_upd** – значение типа bool. Соответствует требованию принудительно обновить данные в базе SQL. Значение по умолчанию False – не обновлять данные принудительно.

Результатом работы функции является запись данных в базу опосредовано через вызов функции query\_splitter() либо обращение к функции load\_data() для выгрузки данных из базы.

## get\_periods(id)

Загружает временные периоды, доступные для выбранного пользователем показателя на сайте ЕМИСС.

Аргументы:

**id** – строковое значение. Номер выгружаемой базы данных по кодировкам ЕМИСС

Возвращает:

**set(periods)** – множество, элементами которого являются кортежи, содержащие уникальные временные периоды показателя.

Функция необходима для последующего сравнения периодов показателя, размещенных ЕМИСС, и периодов, уже загруженных в базу данных.

## monetary\_value()

Составляет отчет о рыночной стоимости нового жилья в многоквартирных домах за 2019-2021 гг.

Выгружает в виде .csv файла.

## monthly\_introduction()

Составляет отчет о ежемесячном вводе нового жилья различных типов в тысячах квадратных метров за 2019-2021 гг.

Выгружает в виде .csv файла.

## monthly\_prices()

Составляет отчет о средней стоимости квадратного метра в первичном и вторичном жилье за 2019-2021 гг.

Выгружает в виде .csv файла.

## user\_interface()

Запрашивает у пользователя id показателя ЕМИСС, с которым он хочет работать. Вызовом функции get\_data() данные этого показателя затем загружаются в базу данных с сервера ЕМИСС. Если данные показателя уже есть в базе, они извлекаются оттуда.

При нажатии пользователем кнопки Enter программа позволяет ввести id нового показателя и описанный процесс повторяется. Остановка произойдет, когда пользователь нажмет Ctrl+C или любую клавишу с последующим нажатием Enter.

1. Тестирование на показателях, отличных от тех, работа с которыми доступна в пользовательском интерфейсе, не проводилось [↑](#footnote-ref-1)