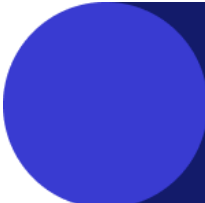


Автоматизация сбора и визуализация данных по показателям целей устойчивого развития (ЦУР) на национальном уровне



Члены команды: Аглиуллина Татьяна, Будко Раиса, Рудаков Максим,
Самойлович Константин, Силиванова Наталья, Чайка Константин

Структура



- Постановка задачи
- Гипотеза
- Цели и задачи
- Модель проекта
- Источники данных
- Используемые методы и подходы
- План реализации исследования
- Решения по визуализации данных
- Результаты

Постановка задачи



- Автоматизированный сбор данных по показателям ЦУР из единой базы ЕМИСС и обеспечение их представление в сводном файле единого формата по причине того, что в настоящее время сбор таких данных осуществляется вручную из электронных таблиц, отличающихся по своему структурному формату



- Автоматизированная система сбора данных по показателям ЦУР, размещенных в ЕМИСС, повысит скорость и качество сбора информации, позволит минимизировать вероятность «ручных» ошибок для последующих публикаций, подготовки отчетов, аналитики и инфографики

Цели и задачи



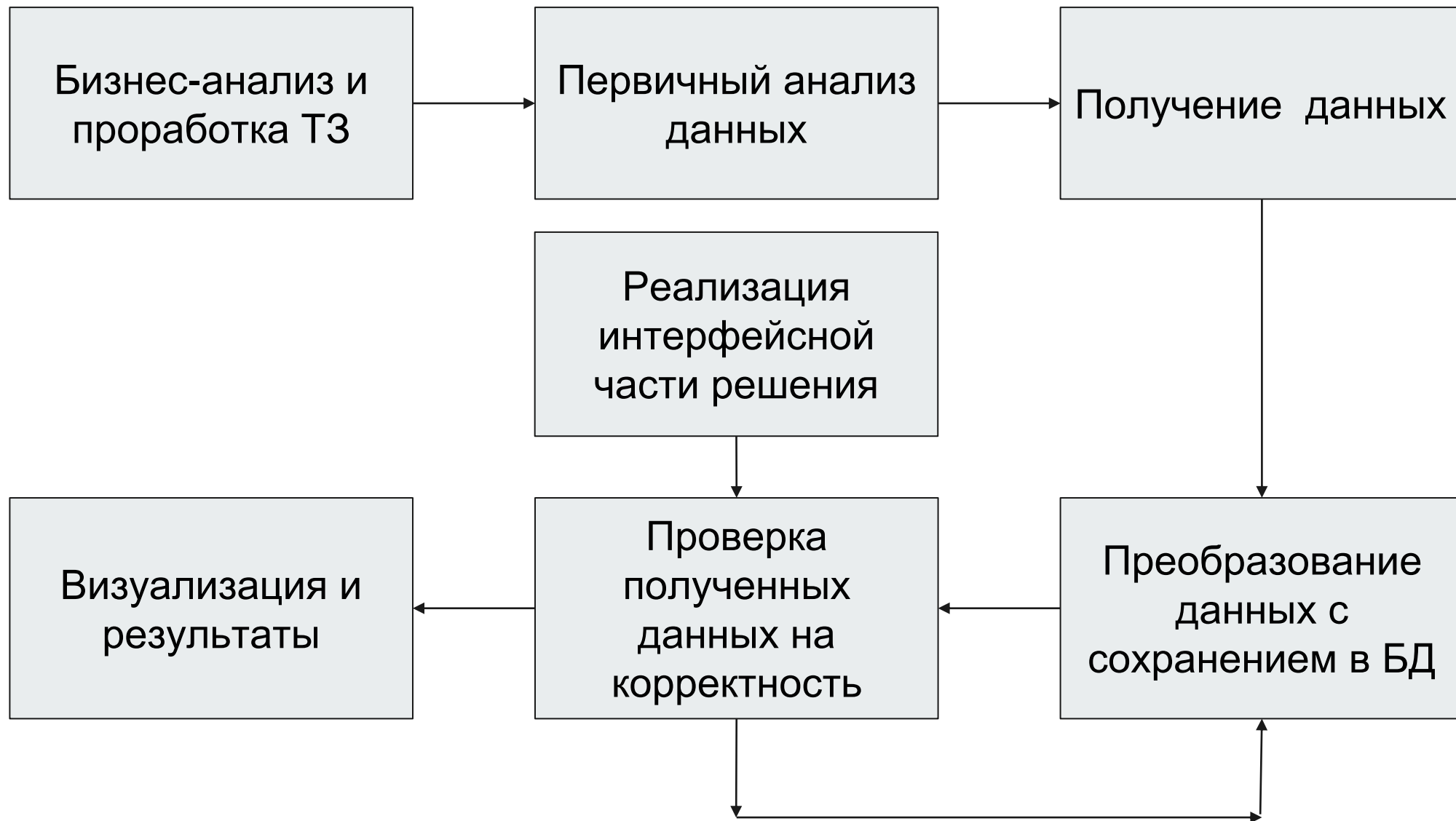
Цель:

- Автоматизация процесса сбора данных по показателям ЦУР, размещенных в ЕМИСС, объединение разрозненных источников данных с большим разнообразием атрибутов и представление их в едином формате

Задачи:

- автоматизация процесса сбора данных по показателям ЦУР
- выбор инструментов для работы с данными
- анализ собранных данных
- объединение данных в сводный документ единого формата
- визуализация полученного набора данных

Модель проекта



Источники данных



Ссылки на открытые источники данных:



ЕМИСС

<https://www.fedstat.ru/>



Федеральная служба государственной статистики

<https://rosstat.gov.ru/sdg/data>



Реестр программного обеспечения

<https://reestr.digital.gov.ru/>

Используемые методы и подходы



- Фреймворк Selenium (Selenoid) - инструмент для автоматизации действий веб-браузера
- BeautifulSoup - парсер для синтаксического разбора файлов
- База данных PostgreSQL - инструмент для записи, хранения, редактирования, удаления данных
- Фреймворк Django на Python - инструмент для создания интерфейса программы
- Celery - способ организации отказоустойчивой очереди получения файлов
- Docker - способ упаковать приложение и все его зависимости в единый образ
- Yandex DataLens - сервис для визуализации данных

План реализации исследования



Этап 1. Получение ссылок на данные показателей ЦУР и скачивание SDMX документов

- 1.1 Разработка парсера для получения доступных ссылок на данные показателей ЦУР в ЕМИСС
- 1.2 Разработка бота на Python с использованием библиотеки Selenium, который пройдет по полученным ссылкам и скачает SDMX из ЕМИСС
 - 1.2.1 Постановка фильтра “выбрать все” для получения всех доступных в системе данных
 - 1.2.2 Снятие предустановленных группировок данных

План реализации исследования



Ограничения этапа 1

- сбор данных осуществляется только по тем показателям, для которых в ЕМИСС содержатся данные
- часть показателей не обладает прямой ссылкой на ЕМИСС, а переходит на промежуточную страницу с выбором классификатора ОКВЭД (поэтому для корректной обработки таких показателей требуется пояснение от Заказчика, какие данные необходимо получить в таком случае, возможность ввода требуемой ссылки реализована в интерфейсе системы)
- для получения данных по некоторым показателям в ЕМИСС требуется авторизация
- в документ SDMX не попадают срезы данных по категориям, которые были предустановлены в поле группировок



EMCC

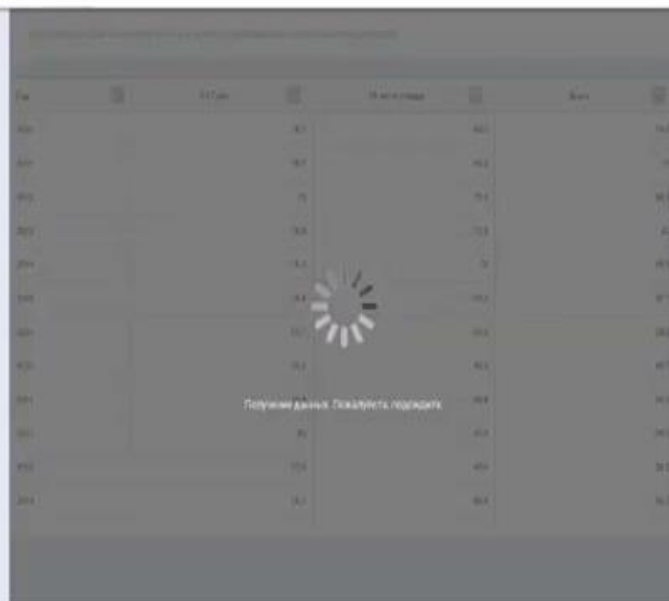
x +

feridat.ru/indicator/158512

🏠 👤 ⓘ

Chrome is being controlled by automated test software.

X



ЕМСС

Информация о сайте

Политика конфиденциальности
Сведения об обработке
Сведения о сайте

Горячая линия

☎ 8 (800) 303-66-47

Электронная почта

✉ info@emss.ru

План реализации исследования



Этап 2. Создание базы данных и парсинг загруженных SDMX документов

- 2.1 Создание структуры базы данных в PostgreSQL
- 2.2 Разработка интерфейса Django для наглядного представления таблиц базы данных
- 2.3 Запись в базу данных полученных SDMX документов в текстовом формате
- 2.4 Парсинг загруженных SDMX документов
- 2.5 Запись в базу данных распарсенных показателей
- 2.6 Создание таблиц словарей в базе данных для хранения уникальных комбинаций “концепт - значение” из кодлиста SDMX документа
- 2.7 Создание кратких обозначений для определенных категорий SDMX документа

План реализации исследования



Ограничения этапа 2

- невалидная структура SDMX документа в некоторых показателях (использование знаков <> в подобной структуре требует специальной обработки, иначе система воспринимает их как теги), которая препятствует корректному сбору данных
- отсутствие унифицированного подхода при создании атрибутов и категорий для показателей на сайте ЕМИСС (в одних показателях категория называется «возраст», в других «разделение по возрасту» и т.д.)

План реализации исследования



Этап 3. Выгрузка данных в едином формате, упаковка web-приложения

- 3.1 Настройка дополнительных опций в интерфейсе (статусы загрузки SDMX документов, статусы активности документов для парсинга и др.)
- 3.2 Настройка выгрузки распарсенных показателей в формат xlsx, обработка данных с помощью библиотеки Pandas
- 3.3 Создание отказоустойчивой очереди опросов получения SDMX документов
- 3.4 Упаковка web-приложения в Docker - контейнер

План реализации исследования



Ограничения этапа 3



Нестабильная скорость работы сайта ЕМИСС в рабочее время, отсутствие доступа к сайту во время проведения технических работ обусловили создание очереди опросов получения SDMX документов (например, скачивание файлов SDMX с определенной периодичностью). То есть пользователю не нужно самостоятельно запускать процессы получения SDMX документов и парсинга их содержимого, система сделает это автоматически. Опциональная настройка этого процесса требует согласования с Заказчиком

План реализации исследования



Этап 4. Визуализация полученного набора данных и оформление результатов



4.1 Загрузка полученного сводного документа по показателям ЦУР на сервис Yandex DataLens, предобработка данных



4.2 Создание дашбордов для визуализации набора данных

Решения по визуализации данных



Ссылки на дашборды:

 <https://datalens.yandex/jouqf3r697v0c?tab=gA>

 <https://datalens.yandex/rvbmks59j5ack>

 <https://datalens.yandex/bg0ozvvtkep04>

Результаты



- Разработано web-приложение для автоматизации сбора данных показателей ЦУР в формате SDMX XML на языке Python с хранением данных в PostgreSQL и выгрузкой в единый файл формата xlsx, для удобства использования приложение обернуто в Docker-контейнер
- Получен сводный документ (xlsx) согласно заданному шаблону по 75 показателям ЦУР (собрано 95 ссылок, часть их них переходят на страницу выбора ОКБЭД, часть имеют требование авторизации. Успешно скачано 77 файлов SDMX, 2 не валидны, итого собраны данные по 75 показателям из 17 групп ЦУР). Итоговый документ содержит 4756 строк и 37 столбцов (каждая вариативная категория вынесена в отдельный столбец с сохранением наименования из кодлиста SDMX документа)



- Web-приложение и сопроводительная документация:
[Ссылка на web-приложение в GitHub](#)
[Пояснения к интерфейсу](#)
- Представлены решения по визуализации набора данных на основе сводного документа показателей ЦУР по следующим категориям: ликвидация нищеты; ликвидация голода; хорошее здоровье и благополучие; недорогостоящая и чистая энергия; достойная работа и экономический рост; индустриализация, инновации и инфраструктура; уменьшение неравенства; сохранение экосистем суши; партнерство в интересах устойчивого развития

Результаты



Корректность полученных данных достигнута двумя путями:

1. Выбором инструментов сбора и чтения данных с реализацией возможности отслеживания процесса на всех этапах получения и анализа данных
2. Сравнением собранных данных с данными на странице ЕМИСС вручную. Выявлено различие по некоторым показателям, обусловленное наличием срезов данных на сайте ЕМИСС



Скорость сбора данных достигнута автоматизацией процесса скачивания. При наличии устойчивого соединения с сайтом скачивание всех показателей с установкой всех фильтров и сбросом группировок занимает 1 час, т.к. днем сайт ЕМИСС имеет сложности с доступом, предлагаем опционально настроить отказоустойчивую очередь опросов получения SDMX документов (например, скачивание файлов SDMX с определенной периодичностью)

Результаты



- Функциональность массива данных достигнута приведением данных к единому формату согласно заданному шаблону. Категории, отсутствующие в шаблоне, вынесены в отдельные столбцы; номера показателей также вынесены в отдельные столбцы

Ссылка на итоговый файл `xlsx`: [Сводная таблица по всем показателям](#)

- Информативность и качество визуализации достигнуты посредством сервиса Yandex DataLens с использованием следующих чартов: диаграмма (круговая, древовидная, столбчатая, точечная, линейчатая, линейная, кольцевая, диаграмма с областями); сводная таблица и карта

Над задачей работали:



группа по разработке технического решения:

Аглиуллина Татьяна

Будко Раиса



группа по визуализации и представлению результатов:

Рудаков Максим

Самойлович Константин

Силиванова Наталья

Чайка Константин