# Постановка задачи

Разработать ПО, с помощью которого будет осуществляться обработка реестра лекарственных средств и медицинских изделий, взятого с конкретного сайта. В нашем случае данным сайтом является государственный реестр лекарственных средств и медицинских изделий Казахстана <http://register.ndda.kz/register.php/mainpage/reestr/lang/ru>. Обработка осуществляется посредством парсинга.

У разрабатываемой нами программы (далее «парсер») есть ряд требований, которых нам необходимо придерживаться:

* Должен осуществляться парсинг всего перечня лекарственных средств и медицинских изделий, присутствующих в реестре
* Сохранение данных должно осуществляться в файл формата .json (по требованию заказчика) на сервер, с которого запускается скрипт
* Дубликаты данных по регистрационному номеру в итоговом файле должны отсутствовать
* Парсер должен предусматривать возможность регулярного запуска, и иметь возможность отслеживания успешности его работы

Парсинг — процесс сбора и систематизации информации, размещенной на определенных сайтах, с помощью специальных программ, автоматизирующих процесс.

# Выбор инструментов реализации

Для решения данной нам задачи требуется определить перечень программных инструментов и технологий, которые мы будем использовать в дальнейшей работе.

Первостепенной задачей является выбор оптимального языка программирования, с помощью которого будет вестись разработка данной программы. Нужно учитывать специфику задачи и поставленные перед нами критерии её выполнения.

В нашем случае оптимальный язык программирования должен обладать:

* удобными инструментами для работы с сетевыми запросами и операциями
* возможностью быстрой обработки больших объемов данных
* универсальностью для выполнения на всех возможных платформах, а также дальнейшего дополнения программы в разных направлениях

Наиболее подходящим языком программирования по всем вышеперечисленным критериям является python так как:

* у него есть широкий спектр различных встроенных и сторонних библиотек, среди которых есть те, которые хорошо подходят под наши задачи
* он обладает подробной, удобной документацией и отзывчивым сообществом
* он является интерпретируемым языком программирования, что позволяет ему достичь максимальной кроссплатформенности и универсальности

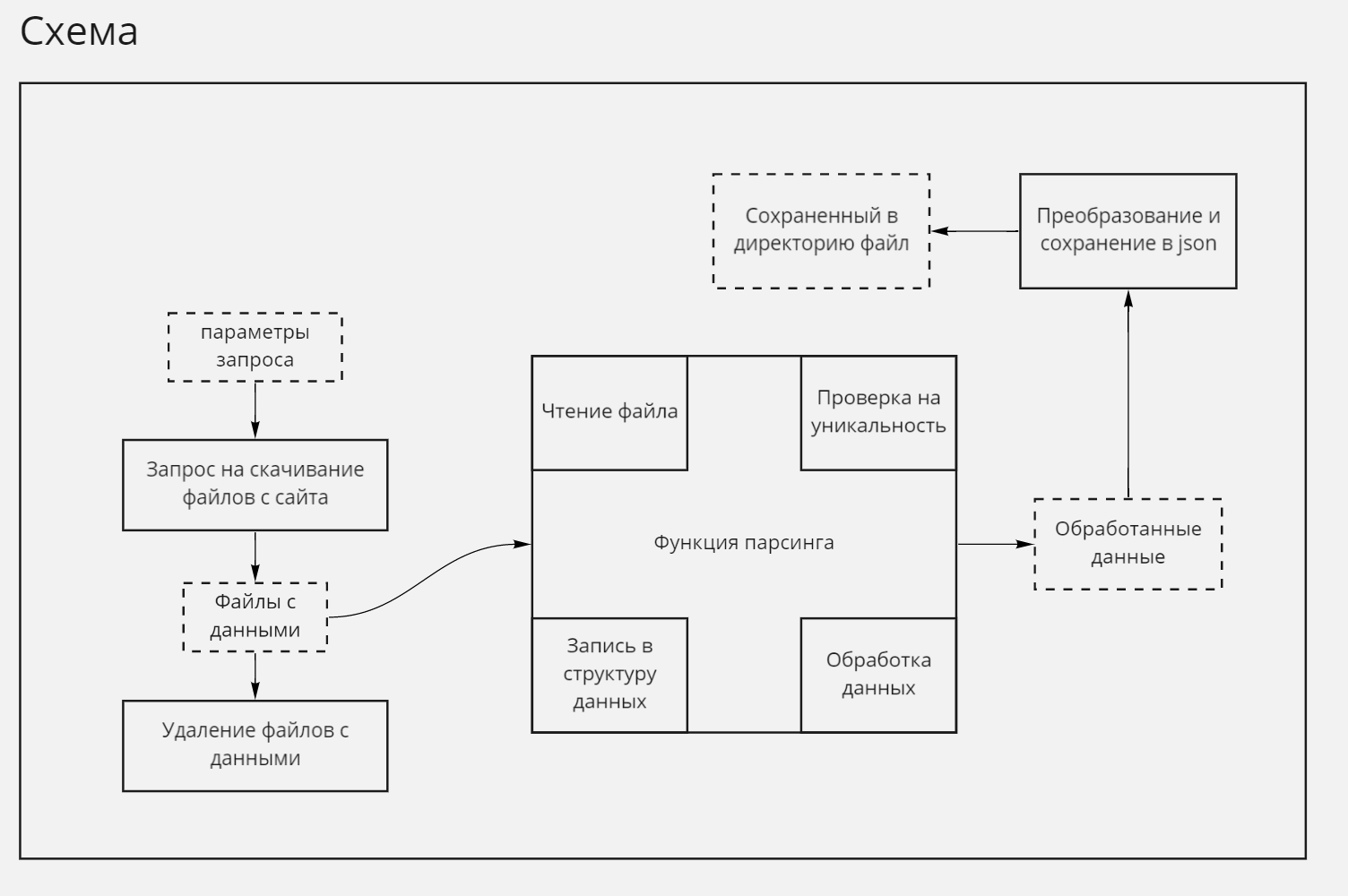
Выбор библиотек исходил из поставленной задачи и вышеперечисленных критериев:

* Для того чтобы получить данные, необходимые для работы, необходимо отправить HTTP запросы к серверу. Для этого была выбрана библиотека requests, как наиболее удобный инструмент.
* Данные, которые получаются с сервера приходят в формате .xls. Для чтения и обработки данного формата была выбрана библиотека xlrd как наиболее компактный и наименее требовательный к ресурсам инструмент.
* Для того чтобы работать с файлами формата json, используется встроенная библиотека json.
* Так как в тексте есть элементы верстки, для обработки текстовых данных при парсинге была выбрана встроенная библиотека html.
* Для вывода времени работы используется библиотека time.
* Библиотека os имеет очень много полезных функций, но в нашем случае используется чтобы очистить данные в консоли и удалить загруженные файлы с данными.
* Для оптимизации процесса работы программы было решено использовать многопоточность процессов. Это позволяет реализовать встроенный модуль multiprocessing.
* Как альтернативный вариант загрузки файлов была предусмотрена работа напрямую с браузером. Для этого используется библиотека selenium.

# Проектирование приложения

Исходя из формулировки задачи и перечисленных выше требований заказчиков к ней, программа должна обладать следующей функциональностью:

1. Получение данных с сайта в формате xls
2. Извлечение данных из файлов
3. Обработка данных
4. Запись данных в удобную структуру данных
5. Преобразование структуры данных в файл формата json (по требованию заказчика)
6. Сохранение файла в директорию



Описание схемы:

Для того, чтобы отправить запрос на скачивание необходимых программе файлов, нужно указать его параметры. На выход этой функции мы получаем файлы с данными, которые идут в функцию парсинга. Там считываются файлы и данные из них обрабатываются и записываются в определенную структуру данных, при этом, идёт проверка на уникальность каждого лекарственного средства и медицинского изделия. На выходе получаются обработанные данные, расположенные в структуре данных, которые преобразовываются в файл формата json и этот файл сохраняется в директорию. В последнюю очередь, удаляются загруженные файлы с данными, так как они были нужны только для того, чтобы данные внутри них были обработаны и преобразованы в json формат. Они больше не нужны и только занимают место.