# 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\*. Обработка осуществляется посредством парсинга.

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

У разрабатываемой нами программы (далее «парсер») есть ряд критериев, 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 на сервер, с которого запускается скрипт
* Дубликаты данных в итоговом файле должны отсутствовать
* Парсер должен предусматривать возможность регулярного запуска, и иметь возможность отслеживания успешности его работы

\* <http://register.ndda.kz/register.php/mainpage/reestr/lang/ru>

# Выбор средств

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 иметь удобные инструменты для работы с сетевыми запросами и операциями, для быстрой обработки больших объемов данных и 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. Также дополнительными плюсами данного языка будет то, что он обладает подробной, удобной документацией и отзывчивым сообществом. Python 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.
* 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 и схема



