# Introduction

## Purpose

В данном документе описаны требования к программному продукту, специализирующемуся в области химии и производящему сравнение химических патентов, их кластеризацию и выявление патентных трендов с учётом упоминаемых в них химических формул.

## Scope

Приложение «Анализатор химических патентов» – это веб-приложение, которое помогает эксперту патентного ведомства затрачивать меньше времени на поиск схожих химических патентов при рассмотрении заявки, а также проводить автоматизированную кластеризацию и поиск схожих патентов.

Приложение будет распространяться по корпоративной лицензии для каждого патентного ведомства, приобретающего данный продукт.

Информация об обработанных химических патентах будет содержаться на внутренних накопителях, объединённых в единое хранилище, расположенное на территории патентного ведомства. Обработка и пополнение базы должно осуществляться средствами вычислительного комплекса патентного ведомства.

Со стороны администратора приложение имеет возможность запуска обработки предварительно загруженных в систему патентных архивов.

Со стороны пользователя приложение имеет возможность просмотра информации об обработанных патентах, химических соединениях, патентных трендах, поиска схожих соединений и патентов.

## Definitions, acronyms, and abbreviations

|  |  |
| --- | --- |
| **Термин** | **Определение** |
| Пользователь | Человек, взаимодействующий с системой через веб-приложение. |
| Администратор | Человек, имеющий доступ к управлению системой. |
| Патентное ведомство | Агентство, выдающее патенты изобретателям и компаниям на их продукты и изобретения. |
| Эксперт патентного ведомства | Работник патентного ведомства, занимающийся выдачей заключений по патентам. |

## References

[1] MDL Information Systems, Inc. CTFile Formats / MDL Information Systems, Inc. – San Leandro : MDL Information Systems, 2003. – 106 с.

[2] ChemSpider reaches 50 million compounds URL: http://www.rsc.org/journals-books-databases/librarians-information/librarians-notes/all-articles/2016/jun/chemspider-reaches-50-million-compounds/

[3] NCI/CADD Chemical Resolver – Chemical Identifier Resolver documentation URL : https://cactus.nci.nih.gov/chemical/structure\_documentation

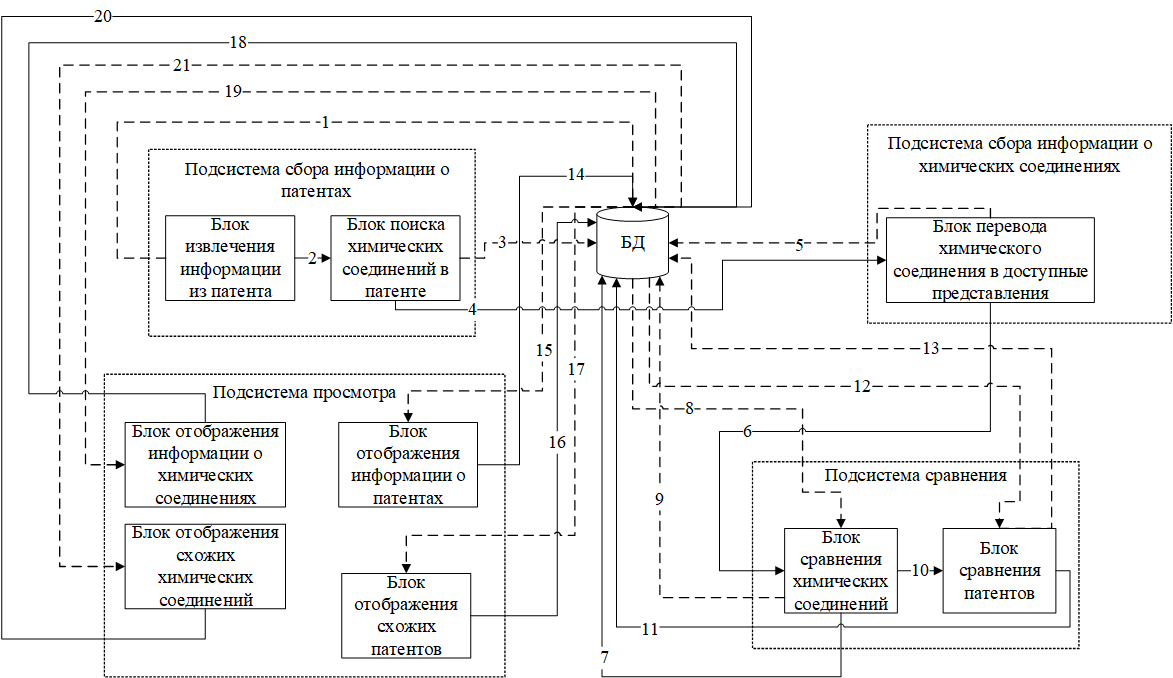
[4] 27 Spring Boot and Spring JDBC with H2 – DZone Java URL: https://dzone.com/articles/spring-boot-and-spring-jdbc-with-h2

## Overview

# Overall Description

## Product Perspective

Данная система будет разработана согласно архитектуре MVC. В качестве модели должна выступать база данных H2, в которую будет вноситься информация об обработанных патентах и химических соединениях. Обработка будет осуществляться в части контроллеров. Общая архитектура разрабатываемого приложения представлена на рисунке 1.

****

1 – сохранение в БД записи патента;

2 – вызов блока поиска химических соединений в патенте;

3 – сохранение в БД записи с информацией о том, что текущие химические соединения содержатся в патенте;

4 – вызов блока перевода химического соединения в доступные представления;

5 – сохранение в БД записи химического соединения;

6 – вызов блока сравнения химических соединений;

7 – запрос к БД на извлечение записей химических соединений;

8 – получение из БД записей химических соединений;

9 – сохранение результатов сравнения химических соединений в БД;

10 – вызов блока сравнения патентов;

11 – запрос к БД на извлечение записей патентов;

12 – получение из БД записей патентов;

13 – сохранение результатов сравнения патентов в БД;

14 – запрос к БД на извлечение записей патентов;

15 – получение из БД записей патентов;

16 – запрос к БД на извлечение записей схожих патентов;

17 – получение из БД записей схожих патентов;

18 – запрос к БД на извлечение записей химических соединений;

19 – получение из БД записей химических соединений;

20 – запрос к БД на извлечение записей схожих химических соединений;

21 – получение из БД записей схожих химических соединений.

Рисунок 1 – Архитектура разрабатываемого приложения

## Product Functions

Данный продукт будет производить обработку химических патентов и соединений, содержащихся в этих патентах. На основе полученных данных будет производиться выдача схожих патентов и химических соединений по запросам пользователя, а также кластеризация патентов и выявление патентных трендов.

## User Classes and Characteristics

Для данного продукта существует два типа пользователей, взаимодействующих с системой: пользователь (эксперт патентного ведомства) и администратор.

Пользователи могут взаимодействовать только с front-end частью приложения, просматривая базу данных патентов и химических соединений, результаты выявления патентных трендов, а также посылая запросы на поиск схожих патентов и химических соединений.

Администратор помимо функционала пользователя может взаимодействовать с back-end частью, загружая в систему патентные архивы, требующие обработки.

## Operating Environment

Продукт может работать под управлением операционных систем семейства Ubuntu (16 LTS и позднее), а также под управлением операционных систем Windows (7 и позднее). Кроссплатформенность обеспечивается за счёт работы системы в Java Run Environment. Для сборки и развёртывания приложения должен использоваться Gradle. Для работы с системой со стороны пользователя необходим веб-браузер.

## User Documentation

Вместе с продуктом будет поставляться документация, включающая в себя руководство оператора и руководство администратора.

## Assumptions and Dependencies

Предполагается, что к определённому этапу реализации проекта, в частности, к началу тестирования отдельных модулей, будет доступен массив патентных архивов USPTO.

Предполагается, что пользователь знаком с основами работы с персональным компьютером и сопутствующей ему периферией, а также умеет работать с интернет-браузером.

Так как данная система будет представлять собой веб-приложение, необходимо использовать интернет-браузер. Предполагается, что пользователи будут иметь бесперебойное подключение к локальной сети патентного ведомства.

# External Interface Requirements

## User Interfaces

Макет страницы с информацией о выбранном химическом соединении представлен на рисунке 2.

Макет страницы с информацией о выбранном патенте представлен на рисунке 3.

Макет страницы со списком химических соединений, схожих с выбранным, представлен на рисунке 4.

Макет страницы со списком патентов, схожих с выбранным, представлен на рисунке 5.

Макет страницы с патентным ландшафтом (патентные тренды) представлен на рисунке 6.

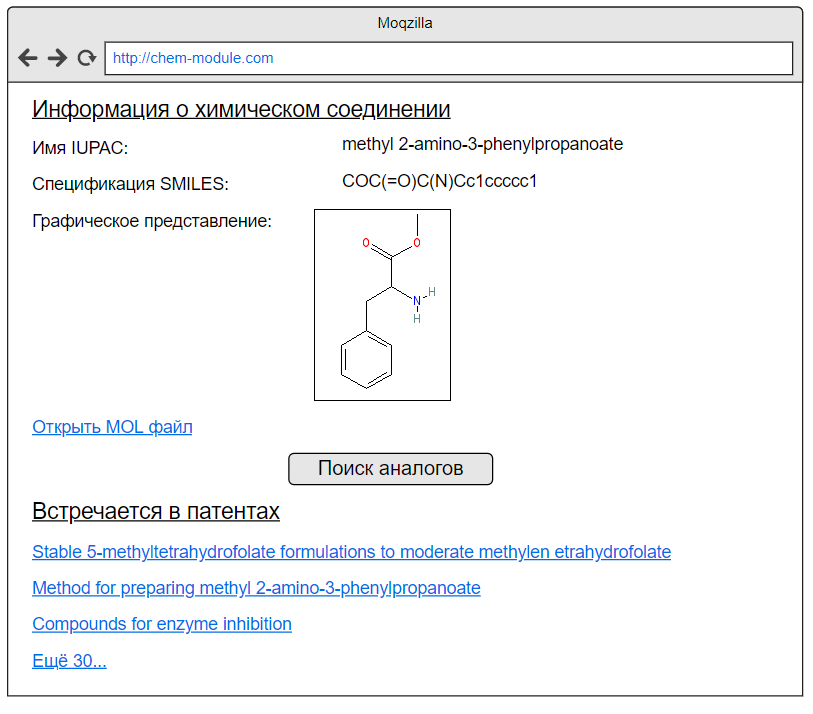


Рисунок 2 – Страница с информацией о выбранном химическом соединении

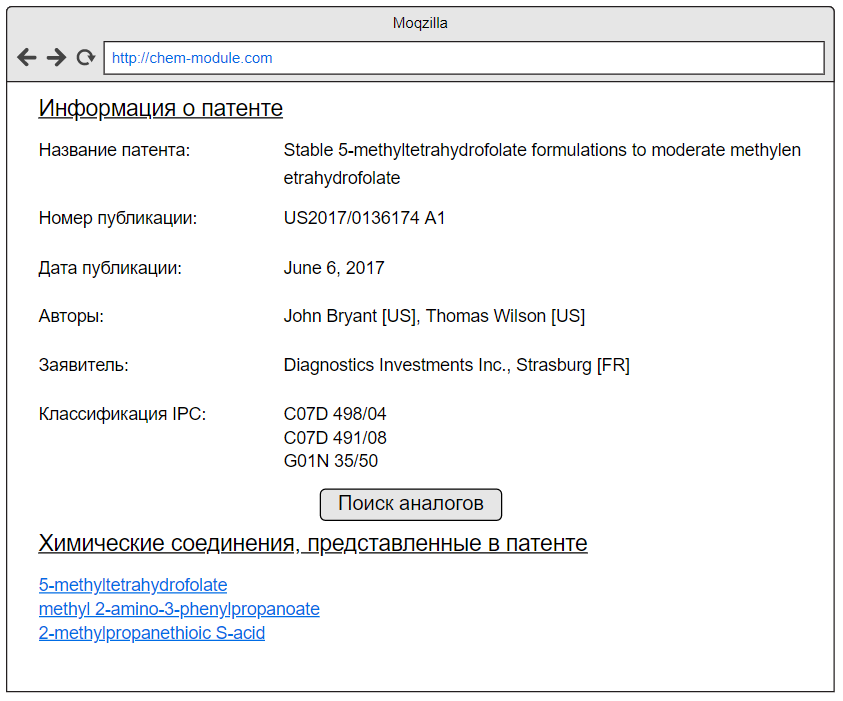


Рисунок 3 – Страница с информацией о выбранном патенте

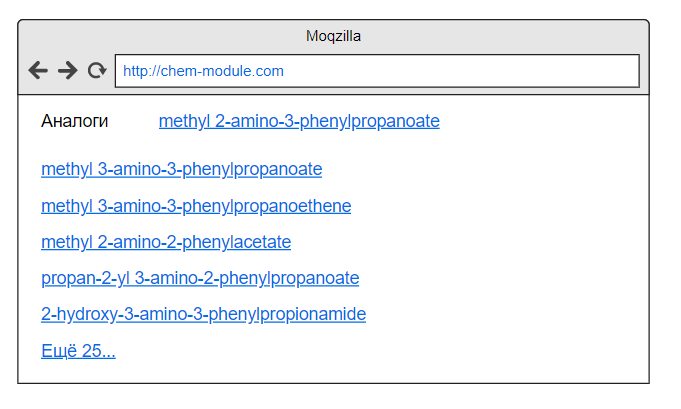


Рисунок 4 – Страница со списком химических соединений, схожих с выбранным

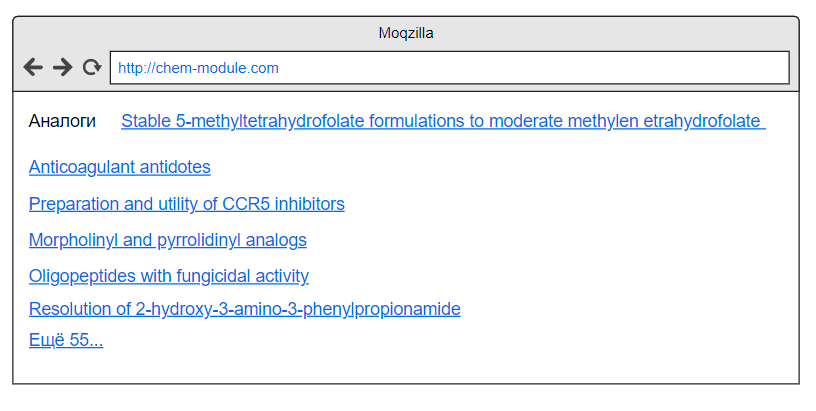


Рисунок 5 – Страница со списком патентов, схожих с выбранным

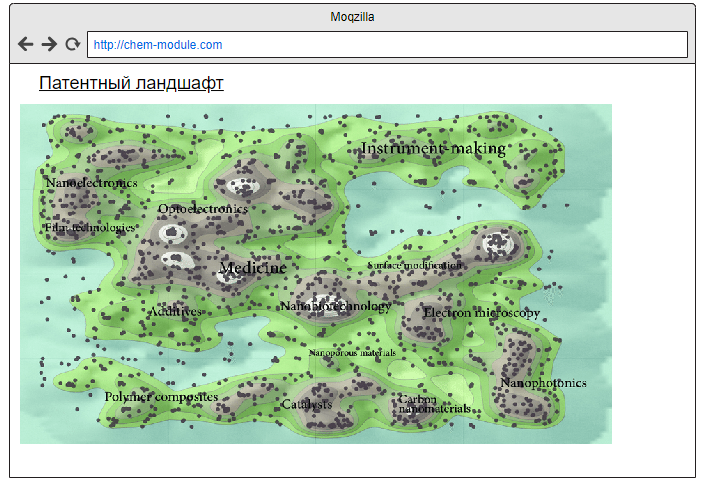


Рисунок 6 – Страница с патентным ландшафтом (патентные тренды)

## Hardware Interfaces

Невозможно обозначить какие-либо конкретные аппаратные решения для разрабатываемой системы, так как она может быть развернута как на персональном компьютере, так и на высокопроизводительном вычислительном кластере. От параметров производительности системы зависит общее время обработки патентных архивов.

## Software Interfaces

Веб-приложение будет работать с помощью фреймворка Spring Boot. Для получения, обработки и записи данных будет использоваться СУБД H2, работающая по TCP.

## Communications Interfaces

Так как приложение монолитно и будет работать только в локальной сети, нет необходимости в задействовании каких-либо протоколов взаимодействия.

# System Features

## Распаковка патентных архивов

Для дальнейшей обработки патентов необходимо извлечь их из представленных .tar архивов.

Извлечение производится в общую папку и извлекаются только патенты химических классов.

Важность: высокая.

## Извлечение информации о патенте

Выбранный патент проходит парсинг. В базу данных производится запись следующей информации о патенте:

* дата,
* название,
* авторы,
* заявитель,
* категория по Международной Патентной Классификации.

Приоритет: средний.

## Извлечение информации о химическом соединении

Из рассматриваемого патента извлекаются химические формулы. В базу данных производится запись следующей информации о химическом соединении:

* графическое представление формулы,
* SMILES-представление формулы,
* InChI-представление формулы,
* ИЮПАК-представление формулы,
* файл химической таблицы.

Приоритет: высокий.

## Поиск схожих соединений

Для выбранной пары химических соединений производится вычисление их молекулярных отпечатков. Затем производится сравнение полученных отпечатков по коэффициенту Танимото. Коэффициент отражает сходство химических соединений.

Приоритет: высокий.

## Поиск схожих патентов

Для выбранной пары патентов производится попарное сравнение имеющихся в них молекулярных отпечатков. Самый большой коэффициент сходства химических соединений принимается за коэффициент сходства патентов.

Приоритет: средний.

## 4.6. Просмотр информации о патентах и схожих патентах, химических соединениях и схожих химических соединениях пользователем

Извлеченная и записанная в базу данных информация может быть просмотрена пользователем через веб-интерфейс приложения.

Приоритет: низкий.

**4.7. Кластеризация патентов по различным критериям**

Массив извлечённых патентов разбивается по подклассам и образует кластеры.

Приоритет: высокий.

**4.8. Выявление патентных трендов**

Для выявления патентных трендов отслеживаются изменения патентных кластеров во времени.

Приоритет: средний.

# Other Nonfunctional Requirements

## Performance Requirements

Минимальные системные требования:

* операционная система семейства Ubuntu (16 LTS или младше), либо Windows (7 или младше);
* 4 Гб ОЗУ;
* процессор Intel i5 или младше;
* 20 Гб свободной памяти на HDD.

## Security Requirements

Во избежание несанкционированного доступа данный программный продукт должен использоваться только внутри локальной сети патентного ведомства.