**Министерство образования и науки Российской Федерации**

Федеральное государственное бюджетное образовательное учреждение высшего образования

**«Владимирский государственный университет**

**имени Александра Григорьевича и Николая Григорьевича Столетовых»**

**«ВлГУ»**

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

Студент

Институт

Направление

**Тема выпускной квалификационной работы**

Руководитель ВКР

(подпись) (ФИО)

Студент

(подпись) (ФИО)

**Допустить выпускную квалификационную работу к защите**

**в государственной экзаменационной комиссии**

Заведующий кафедрой

(подпись) (ФИО)

« » 20 г.

Лист задания

Аннотация на русском

Аннотация на английском

Оглавление

[Введение 3](#_Toc514406552)

[1. Постановка задачи 5](#_Toc514406553)

[1.1 Анализ объекта автоматизации 5](#_Toc514406554)

[1.2 Анализ существующих решений 10](#_Toc514406555)

[1.2.1 MDM-системы 10](#_Toc514406556)

[1.2.2 Talend Open Studio 13](#_Toc514406557)

[1.2.3 Pentaho Data Integration 13](#_Toc514406558)

[1.2.4 Сторонние сервисы 15](#_Toc514406559)

[1.2.5 Исследовательские системы 16](#_Toc514406560)

[1.2.6 Dedupe 16](#_Toc514406561)

[1.2.7 Duke 20](#_Toc514406562)

[1.2.8 Python Record Linkage Toolkit 22](#_Toc514406563)

[1.2.9 Сравнительный анализ библиотек 22](#_Toc514406564)

[2. Проектирование системы 26](#_Toc514406565)

[2.1 Функциональные требования 26](#_Toc514406566)

[2.2 Нефункциональные требования 27](#_Toc514406567)

[2.3 Структура базы данных 28](#_Toc514406568)

[2.4 Алгоритм работы приложения 28](#_Toc514406569)

[3. Реализация системы 29](#_Toc514406570)

[3.1 Загрузка и предобработка данных 29](#_Toc514406571)

[3.2 Поиск дубликатов 32](#_Toc514406572)

# Введение

Объём информации, который приходится хранить и обрабатывать современным информационным системам, растет экспоненциально. Информация – важный ресурс, позволяющий частным компаниям и государственным учреждениям прогнозировать ситуацию на рынке, повышать свою производительность, улучшать работу предприятия по многим экономическим показателям. Для того, чтобы получить корректный результат работы системы и пользоваться всеми преимуществами владения информацией, необходимо гарантировать её качество. Одна из проблем, с которой приходится сталкиваться при работе с данными, это повторяемость. К огромному сожалению, не существует стандарта, с помощью которого можно было бы однозначно отобразить любую часть окружающего мира в информационный объект. Широкая вариативность описания – одна из причин по которой каждая информационная система имеет свою структуру данных для хранения информации о материальных объектах реального мира. Разночтения могут заключаться, например, в количестве имеющихся атрибутов, наименованиях, формате и даже в типе данных для отображения одного и того же свойства идентичных объектов. Не стоит забывать и о том, что львиная доли информации поступает напрямую от человека и важно контролировать её достоверность и актуальность. В данные могут закрадываться грамматические и синтаксические ошибки, они могут нарочно или случайно дублироваться пользователем, что внесет значительную погрешность в конечный результат. Кроме того, на определенном этапе жизненного цикла может возникнуть необходимость интеграции нескольких информационных систем. Перечисленные выше разночтения значительно затруднят интеграцию данных и поставят под сомнение их качество. Задачей сопоставления данных между собой занимается такое направление как dublicate detection. Одна из задач, решаемых в этой области, это определение связи объекта с реальным миром. Такое сопоставление позволяет избавиться от дубликатв, первопричиной которых может являться как пользователь, так и неграмотно выстроенный процесс интеграции систем. Иронично, что для обозначения этого направления, существует ещё около десятка дублирующих названий: Recode linkage, Dublicate detection, Deduplication, Reference reconciliation, Reference matching, Object consolidation, Fuzzy matching, Entity clustering, Hardening soft databases и т.д.