Проект «Создание цифровой платформы фольклорного архива УрФУ»

# РЕФЕРАТ

- на русском

- на английском

# МЕСТО ВЫПОЛНЕНИЯ РАБОТЫ

# СОДЕРЖАНИЕ

Оглавление

[РЕФЕРАТ 2](#_Toc41165982)

[МЕСТО ВЫПОЛНЕНИЯ РАБОТЫ 3](#_Toc41165983)

[СОДЕРЖАНИЕ 4](#_Toc41165984)

[ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ 5](#_Toc41165985)

[ВВЕДЕНИЕ 6](#_Toc41165986)

[ОСНОВНАЯ ЧАСТЬ 7](#_Toc41165987)

[Обзор литературы 7](#_Toc41165988)

[Постановка задачи работы 8](#_Toc41165989)

[Раздел 9](#_Toc41165990)

[Исследование потребностей 9](#_Toc41165991)

[Roadmap 11](#_Toc41165992)

[Исследование конкурентов 12](#_Toc41165993)

[Сценарии 13](#_Toc41165994)

[MVP 15](#_Toc41165995)

[Проектирование базы данных 16](#_Toc41165996)

[Проектирование API 22](#_Toc41165997)

[ЗАКЛЮЧЕНИЕ 24](#_Toc41165998)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ И ЛИТЕРАТУРЫ 25](#_Toc41165999)

[ПРИЛОЖЕНИЯ 26](#_Toc41166000)

# ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

УрФУ

УрГУ

CMS система

БД

MVP

Гипотеза - предполагаемая проблема

Roadmap, или дорожная карта — это план создания продукта.

# ВВЕДЕНИЕ

Фольклорный архив УрФУ содержит более 250 000 фольклорных текстов. Этот материал собирался на протяжении 60 лет работы фольклорной экспедиции УрГУ/УрФУ. В настоящие время архив существует только в рукописном виде. Это одна из крупнейших региональных фольклорных коллекций в стране.

В настоящий момент в России не существует подобных электронных ресурсов, многие фольклорные материалы не опубликованы, хранятся в университетских депозитариях в бумажном виде и соответственно остаются труднодоступными для исследователей.

Для обеспечения компактного хранения больших объемов документов и организации системы их быстрого поиска нужно создать электронный архив. Данная система позволит массово оцифровать бумажные документы, систематизировать их и надежно хранить большой объём электронных документов. Поисковые возможности, которые могут быть реализованы при создании электронного корпуса русского фольклора, в перспективе позволит совершенно по-новому представить картину бытования фольклора во времени и пространстве, в различных социальных средах, а также по-новому взглянуть на историю полевой фольклористики в нашей стране.

Оцифровка и научная онлайн публикация фольклора, как уже опубликованного в научных изданиях, так и существующего лишь в виде карточек полевых экспедиций, может стать важным имиджевым проектом для УрФУ, поскольку неизбежно привлечет к себе внимание исследователей как в России, так и за рубежом. Масштабирование проекта существенно повышает его национально-культурное и научное значение, а значит, позволит привлекать регулярное грантовое финансирование. Кроме того, проектируемая CMS система в дальнейшем может быть адаптирована для создания аналогичных коллекций литературных и языковых материалов, коллекций исторических источников и т.п.

# ОСНОВНАЯ ЧАСТЬ

## Обзор литературы

## Постановка задачи работы

В рамках проекта требуется:

1. Оцифровать тексты, создать электронные копии текстовых материалов
2. Провести аналитику и выявить основные требования к проекту, выяснить проблемы и сценарии
3. Спроектировать электронную базу данных с указанием для каждого текста атрибутов и классификационных признаков, которые могут быть использованы для работы с БД
4. Спроектировать API архива для работы с документами
5. Разработать электронную онлайн платформу, позволяющую, с одной стороны, удаленно пополнять электронную коллекцию фольклорных текстов, а с другой стороны формировать выборки по заданным параметрам, осуществлять полноценный поиск
6. Сформировать MVP и разработать ядро CMS и заполнить его уже оцифрованными материалами фольклорного архива УрФУ

## Раздел

### Исследование потребностей

Для формирования качественного и нужного продукта нужно провести исследование потребностей клиента. Одна из методик проведения исследования – глубинное интервью.

Глубинное интервью – это беседа с респондентом в форме, побуждающей последнего к подробным ответам на заданную тему. Позволяет выявить сценарии, скрытые мотивы, ценности, убеждения, проблемы и потребности респондента.

Интервью включает в себя:

- Формулировку целей исследования

- Составление списка вопросов

- Проведение интервью

- Обработка результатов

Цель исследования:

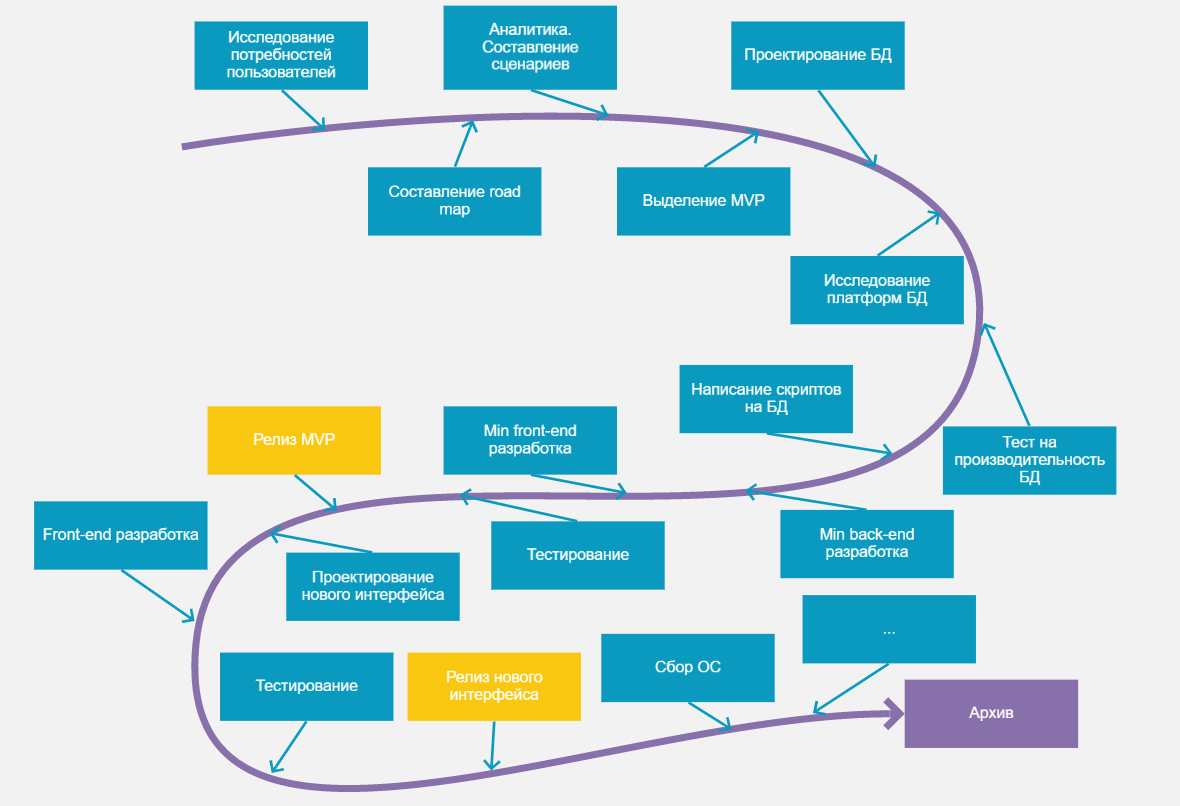
Формирование ценности продукта, исследование сценариев пользователей, выделение активных ролей, формирование основной функциональности, формирование дополнительной функциональности.

Результаты исследования:

|  |  |  |
| --- | --- | --- |
| № | Вопрос | Ответ |
| 1. | Для чего нужен проект? В чем ценность? | Для формирования электронного архива.  Ценность: быстрый доступ |
| 2. | Какой формат документов? | В основном это рукописные записи в очень плохом состоянии. Предполагается, что хранить будут сканы и их расшифровки |
| 2. | Какую дополнительную информацию нужно хранить у документа? | - жанр  - место  - время текста  - информант  - фольклорист  - мотивно-тематическая классификация  - ключевые слова  - источник  - библиография о тексте  - морфологический анализ |
| 3. | Кто пользуется системой? | Фольклористы – исследователи |
| 4. | Как каждая роль использует систему? Какую функцию выполняет? | Есть обычные пользователи. Они используют систему для поиска нужных документов. Поиск происходит по дополнительной информации.  Есть администраторы, которые могут добавлять или удалять документы, редактировать информацию |
| 5. | Без чего на первом этапе проект не может существовать? | Без добавления/удаления/редактирования документов |
| 6. | Как в дальнейшем хочется развивать проект? | Составить карту фольклорного быта |
| 7. | Где разворачивается проект? На серверах УрФУ? | На серверах УрФУ |
| 8. | Какие есть технические ограничения на серверах УрФУ? | Нет ограничений |

### Roadmap

Для успешного запуска проекта требуется составить план. Основа плана — стратегия, а для ее реализации требуется вспомогательное средство. Это roadmap или дорожная карта. Инструмент, который используется в разработке сложных технических продуктов.



Этапы создания продукта:

1. Исследование потребностей
2. Составление roadmap
3. Аналитика, составление сценариев
4. Выделение MVP
5. Проектирование БД
6. Написание скриптов для БД
7. Реализация минимального back-end
8. Проектирование API
9. Реализация минимального front-end
10. Тестирование
11. Релиз

### Исследование конкурентов

В разборе сценариев, портретов клиентов, потребностей пользователей очень помогает исследование конкурентов и поиск аналогичных систем.

Я нашла несколько похожих систем:

1. Национальный корпус русского языка - http://www.ruscorpora.ru.

На этом сайте помещен корпус современного русского языка общим объемом более 600 млн слов. Корпус русского языка — это информационно-справочная система, основанная на собрании русских текстов в электронной форме.

Есть новости сайта. Система не стоит на месте и развивается, а, следовательно, выходят обновления и появляются новости.

Поиск текстов. Есть сложный поиск, есть поиск по регулярным выражениям. Есть также поиск по различным критериям, таких как жанр, грамматический признак и т.д.

1. Национальная электронная библиотека - https://rusneb.ru/

На данный момент в НЭБ доступно 1 671 878 электронных книг и 29 948 245 записей каталогов. Это древние рукописи, советские журналы, учебники и т.д. Оцифрованные документы можно скачать или полистать прямо в браузере.

В разделе «Коллекции» составляются подборки по авторам и событиям.

### Сценарии

В нашей системе есть 2 вида пользователей: администратор и читатель.

Функциональность для читателя:

1. Найти книги по:
   1. жанру
   2. месту
   3. времени текста
   4. информанту
   5. фольклористу
   6. мотивно-тематической классификации
   7. ключевым словам
   8. источнику
   9. библиографии о тексте
   10. любой комбинации из выше стоящих элементов
2. Открыть карточку документа или сам документ
3. Скачать документ
4. Зарегистрироваться в системе и получить доступ к функциональности администратора

Функциональность администратора:

1. Найти книги по:
   1. жанру
   2. месту
   3. времени текста
   4. информанту
   5. фольклористу
   6. мотивно-тематической классификации
   7. ключевым словам
   8. источнику
   9. библиографии о тексте
   10. любой комбинации из выше стоящих элементов
2. Авторизоваться
3. Изменить параметры у документа
   1. жанр
   2. место
   3. время текста
   4. информант
   5. фольклорист
   6. мотивно-тематическая классификация
   7. ключевые слова
   8. источник
   9. библиография о тексте
4. Изменить сам документ
5. Удалить документ
6. Удалить дополнительную информацию у самого документа
7. Добавлять других администраторов

### MVP

Минимально жизнеспособный продукт или minimum viable product — продукт, обладающий минимальными, но достаточными для удовлетворения первых потребителей функциями. Основная задача создания MVP — получение обратной связи для формирования гипотез дальнейшего развития продукта.

### Проектирование базы данных

Проектирования БД – переход от неформального словесного описания информационной структуры предметной области к формализованному описанию объектов предметной области в терминах некоторой модели. Было проведены следующие этапы:

1. Системный анализ и словесное описание информационных объектов предметной области
2. Проектирование базы данных в терминах модели данных
3. Физическое проектирование БД, развертывание в сервисе Azure

Рассмотрим подробнее первый этап – системный анализ и сбор требований. В нашей Базе данных требовалось хранить такие сущности как документ, фольклорист, информант, жанр, классификатор, тег. Рассмотрим подробнее каждую сущность:

Документ – сущность для хранения данных о документе. Например:

1. Идентификатор (Id) – уникальный ключ для документа
2. Название документа (Title) – предполагается, что это будет строка, в которой будет в краткой форме записано название
3. Расшифровка документа (Content) – так как многие тексты в архиве в трудночитаемом состоянии, то многие тексты были переведены в электронный вид. Теперь нудно хранить расшифровку текстов в сервисе
4. Место (PlaceName, PlaceLatitude, PlaceLongitude) – документ имеет принадлежность к месту, где его собрали. Для визуализации текстов на карте требуется кроме названия места хранить долготу и широту
5. Год записи (YearOfRecord) – год, в котором документ был записан в архив
6. Дополнительная информация (AdditionalInformation) – комментарий к документу
7. Файл (FileName, FileId) – в нашем сервисе будет возможность загрузить скан документа в сервис, для этого в сущности документ требуется 2 поля – наименование загруженного файла и идентификатор этого файла для хранения документа на сервере
8. Дополнительные поля для работы сервиса (Deleted, CreatedAt)
   1. Deleted – признак того, что документ был удален. Нужен для показа актуальной информации
   2. CreatedAt – дата создания документа. Нужен для сортировки документов

Скрипт создания документа:

Create Table Documents(

Id INT NOT NULL PRIMARY KEY IDENTITY (1, 1),

Title nvarchar(200),

Content ntext,

PlaceName nvarchar(200),

PlaceLatitude float,

PlaceLongitude float,

YearOfRecord int,

AdditionalInformation ntext,

FileName nvarchar(200),

FileId nvarchar(100),

Deleted bit not null,

CreatedAt date default (getdate())

);

Фольклорист – человек, который сделал запись о документе. В сущность фольклорист нужно указывать следующие параметры:

1. Идентификатор (Id) – идентификатор уникальной записи в таблицу

2. Фамилия Имя и Отчество фольклориста (FIO) – строка для записи ФИО сотрудника

Скрипт для создания таблицы Фольклористы:

Create Table Folklorists(

Id INT NOT NULL PRIMARY KEY IDENTITY (1, 1),

FIO nvarchar(200) NOT NULL);

Информант – человек, который рассказал историю, частушку или другой жанр произведения.

Сущность Информант должна содержать следующие поля:

1. Идентификатор (Id) – идентификатор уникальной записи в таблицу

2. Фамилия Имя и Отчество информанта (FIO) – строка для записи ФИО человека, который предоставил материал

3. Год рождения (YearOfBirth) – год рождения информанта. Нужно для понимания, какие особенности речи характерны для людей определенной эпохи

Скрипт для создания таблицы Информантов:

Create Table Informants(

Id INT NOT NULL PRIMARY KEY IDENTITY (1, 1),

FIO nvarchar(200) NOT NULL,

YearOfBirth INT);

Жанр - совокупность формальных и содержательных особенностей произведения.

Сущность Жанр должна содержать следующие поля:

1. Идентификатор (Id) – идентификатор уникальной записи в таблицу

2. Название жанра (GenreName) – строка, в которой будет хранится название жанра

Скрипт для создания таблицы Жанры

Create Table Genres(

Id INT NOT NULL PRIMARY KEY IDENTITY (1, 1),

GenreName nvarchar(50));

Мотивно-тематическая классификация. Сущность Класификатора должна содержать следующие поля:

1. Идентификатор (Id) – идентификатор уникальной записи в таблицу

2. Код классификации (Code) – строка, в которой будет хранится код

3. Наименование классификации (ClassificationName) – название классификации

Скрипт для создания таблицы Классификации

Create Table MotivationalThematicClassifications(

Id INT NOT NULL PRIMARY KEY IDENTITY (1, 1),

Code nvarchar(10),

ClassificationName nvarchar(100));

Сущность Теги должна содержать следующие поля:

1. Идентификатор (Id) – идентификатор уникальной записи в таблицу

2. Наименование тега (TagName)

Скрипт для создания таблицы Теги

Create Table Tags(

Id INT NOT NULL PRIMARY KEY IDENTITY (1, 1),

TagName nvarchar(50));

Промежуточные таблицы

Для осуществления связей многие ко многим требуется создать промежуточные таблицы для следующих сущностей:

1. Документы – Информанты

2. Документы – Фольклористы

3. Документы – Жанры

4. Документы – Классификации

5. Документы - Теги

Скрипты для создания промежуточных таблиц

Create Table Rels\_Document\_Informant(

Id INT NOT NULL PRIMARY KEY IDENTITY (1, 1),

DocumentId int,

InformantId int

FOREIGN KEY (DocumentId) REFERENCES Documents (Id),

FOREIGN KEY (InformantId) REFERENCES Informants (Id));

Create Table Rels\_Document\_Genre(

Id INT NOT NULL PRIMARY KEY IDENTITY (1, 1),

DocumentId int,

GenreId int

FOREIGN KEY (DocumentId) REFERENCES Documents (Id),

FOREIGN KEY (GenreId) REFERENCES Genres (Id));

Create Table Rels\_Document\_Folklorist(

Id INT NOT NULL PRIMARY KEY IDENTITY (1, 1),

DocumentId int,

FolkloristId int

FOREIGN KEY (DocumentId) REFERENCES Documents (Id),

FOREIGN KEY (FolkloristId) REFERENCES Folklorists (Id));

Create Table Rels\_Document\_Tag(

Id INT NOT NULL PRIMARY KEY IDENTITY (1, 1),

DocumentId int,

TagId int

FOREIGN KEY (DocumentId) REFERENCES Documents (Id),

FOREIGN KEY (TagId) REFERENCES Tags (Id));

Create Table Rels\_Document\_MotivationalThematicClassification(

Id INT NOT NULL PRIMARY KEY IDENTITY (1, 1),

DocumentId int,

MotivationalThematicClassificationId int

FOREIGN KEY (DocumentId) REFERENCES Documents (Id),

FOREIGN KEY (MotivationalThematicClassificationId) REFERENCES MotivationalThematicClassifications (Id));

Также есть отдельная таблица для хранения морфологического анализа.

Create Table Morph(

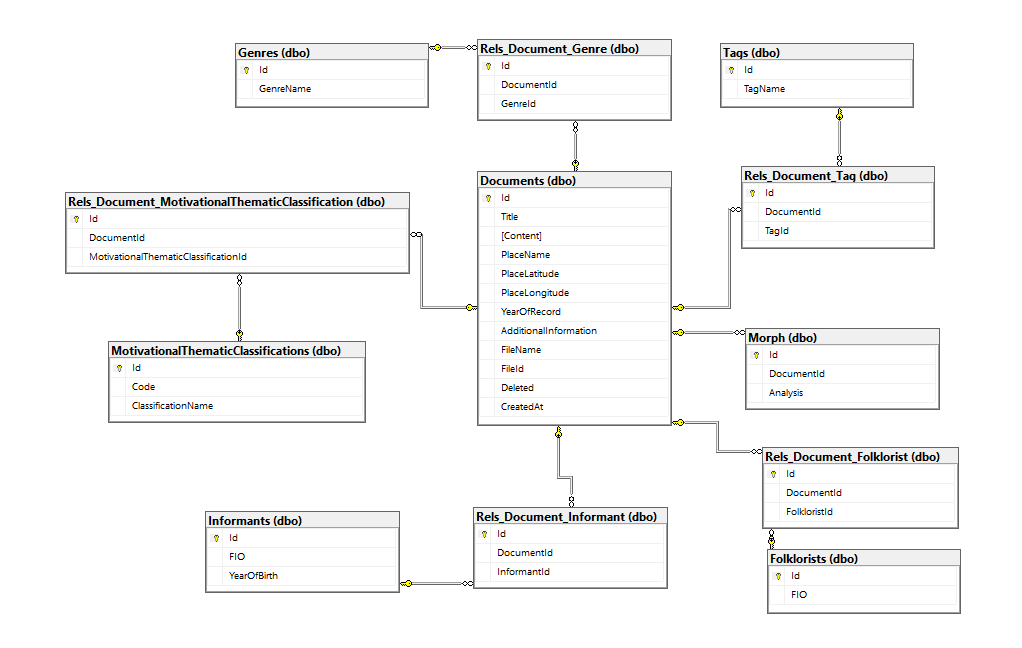
Id INT NOT NULL PRIMARY KEY IDENTITY (1, 1),

DocumentId int,

Analysis ntext

FOREIGN KEY (DocumentId) REFERENCES Documents (Id))

Результат проектирования базы данных представлен на рисунке:



### Проектирование API

Большинство современных веб-приложений предоставляют API, которые клиенты могут использовать для взаимодействия с приложением. Для того, чтобы спроектировать API нужно перечислить все типы данных, которые клиентское приложение может захотеть получить или передать с помощью сервиса. Такое описание называется семантическим, так как отображает значение данных в приложении, и описание того, что происходит в самом приложении.

Сущность «Документ» имеет следующие семантические описания: идентификатор (Id тип int) – уникальный ключ для документа, название документа (Title тип string), расшифровка документа (Content тип string), название места сбора информации и координаты места (PlaceName тип string, PlaceLatitude тип double, PlaceLongitude тип double), год записи документа (YearOfRecord тип int), дополнительная информация (AdditionalInformation тип string), отсканированный файл (FileName тип string, FileId тип string), список жанров (Genres массив типа Genre), список информантов (Informants массив типа Informant), список фольклористов (Folklorists массив типа Folklorist), список мотивов (MotivationalThematicClassifications массив типа MotivationalThematicClassification), список тегов (Tags массив типа Tag) и морфологический разбор текста (Morph тип string).

Сущность «Жанр» имеет следующие поля: идентификатор (Id тип int) – уникальный ключ для жанра, название жанра (GenreName тип string).

У сущности «Информант» есть следующие поля: идентификатор (Id тип int) – уникальный ключ для информанта, фамилия имя и отчество информанта (FIO тип string) и год рождения (YearOfBirth тип int).

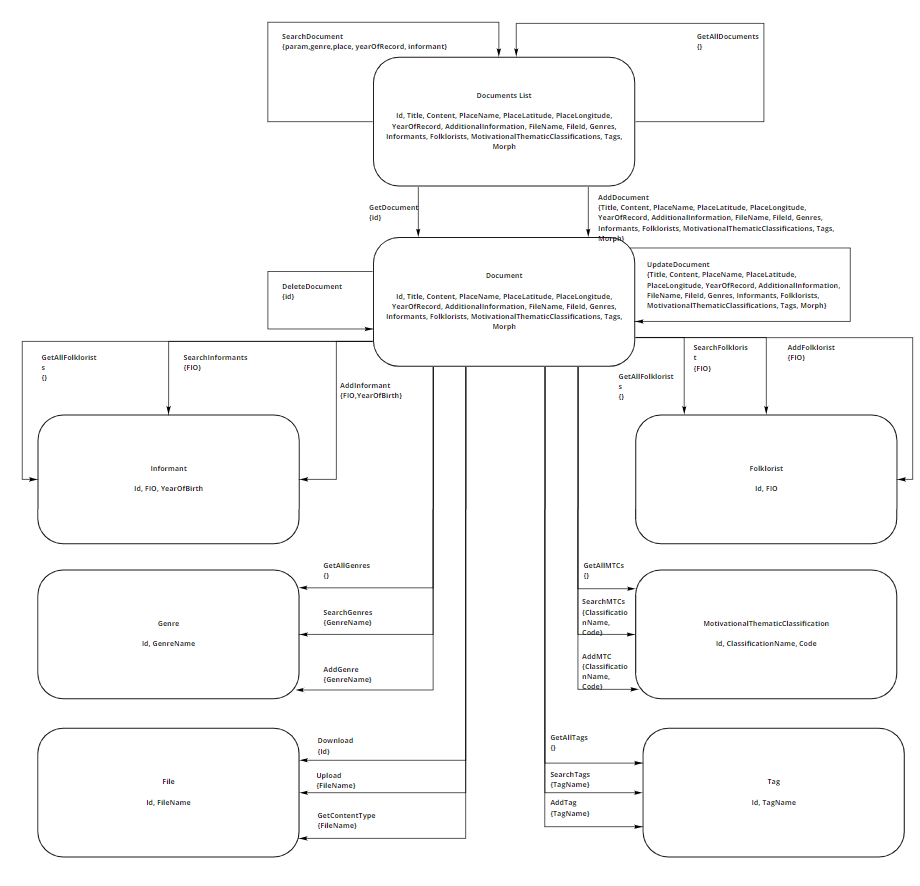
У сущности «Фольклорист» есть следующие поля: идентификатор (Id тип int) – уникальный ключ для фольклориста, фамилия имя и отчество фольклориста (FIO тип string).

Сущность «Мотивно-тематический классификатор» имеет следующие поля: идентификатор (Id тип int) – уникальный ключ для классификатора, название мотива или тематики (ClassificationName тип string) и код (Code тип string)

Сущность «Тег» имеет идентификатор (Id тип int) – уникальный ключ для тега, название тега (TagName тип string).

Следующим этапом спроектируем возможные состояния в сервисе. Для этого используем диаграмму состояний для предполагаемого API. Каждый блок на диаграмме представляет возможное состояние, стрелки используются для обозначения переходов от одного блока к другому, от одного состояния к следующему. Эти переходы инициируются запросами.

Приложение для хранения фольклорных документов может потребовать доступ к карточке документа, возможно фильтровать документы, можно создавать документы, можно искать, создавать и обновлять такие сущности как фольклорист, информант, жанр, мотивно-тематический классификатор, тег. Многие из этих действий используют значения состояний для передачи данных между клиентом и сервером. Например, действие «добавить документ» позволяет клиенту передавать значения состояний Title, Content, PlaceName, PlaceLatitude, PlaceLongitude, YearOfRecord, AdditionalInformation, FileName, FileId, Genres, Informants, Folklorists, MotivationalThematicClassifications, Tags и Morph. Ниже на рис показаны основные действия по документам.



Примененные инструменты и технологии Способы и методы решения задачи

# ЗАКЛЮЧЕНИЕ

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ И ЛИТЕРАТУРЫ

* [https://medium.com/9pm-agency/что-делать-когда-надо-провести-интервью-с-потребителями-4ae5f860883b](https://medium.com/9pm-agency/%D1%87%D1%82%D0%BE-%D0%B4%D0%B5%D0%BB%D0%B0%D1%82%D1%8C-%D0%BA%D0%BE%D0%B3%D0%B4%D0%B0-%D0%BD%D0%B0%D0%B4%D0%BE-%D0%BF%D1%80%D0%BE%D0%B2%D0%B5%D1%81%D1%82%D0%B8-%D0%B8%D0%BD%D1%82%D0%B5%D1%80%D0%B2%D1%8C%D1%8E-%D1%81-%D0%BF%D0%BE%D1%82%D1%80%D0%B5%D0%B1%D0%B8%D1%82%D0%B5%D0%BB%D1%8F%D0%BC%D0%B8-4ae5f860883b)
* <http://bgoal.ru/blog/articles/kak-sozdat-produkt-kotoryy-kupyat-metodologiya-customer-development.html>
* <http://2016.profsoux.ru/papers/paper-143>
* <https://neilpatel.com/blog/26-customer-development-resources/>
* <https://vc.ru/flood/41077-kak-delat-menshe-intervyu-provodya-customer-development>
* [https://medium.com/@ratner.svetlana/%D0%B8%D0%BD%D1%82%D0%B5%D1%80%D0%B2%D1%8C%D1%8E-%D1%81-%D0%BA%D0%BB%D0%B8%D0%B5%D0%BD%D1%82%D0%B0%D0%BC%D0%B8-%D1%84%D0%BE%D1%80%D0%BC%D1%83%D0%BB%D0%B8%D1%80%D1%83%D0%B5%D0%BC-%D0%B3%D0%B8%D0%BF%D0%BE%D1%82%D0%B5%D0%B7%D1%8B-a688561fc30e](https://medium.com/@ratner.svetlana/%D0%B8%D0%BD%D1%82%D0%B5%D1%80%D0%B2%D1%8C%D1%8E-%D1%81-%D0%BA%D0%BB%D0%B8%D0%B5%D0%BD%D1%82%D0%B0%D0%BC%D0%B8-%D1%84%D0%BE%D1%80%D0%BC%D1%83%D0%BB%D0%B8%D1%80%D1%83%D0%B5%D0%BC-%D0%B3%D0%B8%D0%BF%D0%BE%D1%82%D0%B5%D0%B7)
* [https://medium.com/designkontur/%D0%BF%D1%80%D0%BE%D0%B1%D0%BB%D0%B5%D0%BC%D0%BD%D0%BE%D0%B5-%D0%B8%D0%BD%D1%82%D0%B5%D1%80%D0%B2%D1%8C%D1%8E-%D0%B8%D0%BB%D0%B8-%D0%BA%D0%B0%D0%BA-%D0%BF%D1%80%D0%BE%D0%B2%D0%B5%D1%80%D0%B8%D1%82%D1%8C-%D1%81%D0%B2%D0%BE%D1%8E-%D0%B8%D0%B4%D0%B5%D1%8E-%D0%BD%D0%B0-%D0%BF%D0%BE%D0%BB%D1%8C%D0%B7%D0%BE%D0%B2%D0%B0%D1%82%D0%B5%D0%BB%D1%8F%D1%85-2cd409498c5c](https://medium.com/designkontur/%D0%BF%D1%80%D0%BE%D0%B1%D0%BB%D0%B5%D0%BC%D0%BD%D0%BE%D0%B5-%D0%B8%D0%BD%D1%82%D0%B5%D1%80%D0%B2%D1%8C%D1%8E-%D0%B8%D0%BB%D0%B8-%D0%BA%D0%B0%D0%BA-%D0%BF%D1%80%D0%BE%D0%B2%D0%B5%D1%80%D0%B8%D1%82%D1%8C-%D1%81%D0%B2%D0%BE%D1)

# ПРИЛОЖЕНИЯ