ВЕСТНИК ЮГОРСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2020 г. Выпуск 4 (59). С. 13–21

DOI: 10.17816/byusu20200413-21
УДК 004.6:39

В. В. Бурлуцкий, В. А. Балуев, М. И. Изерт, М. А. Русанов

РЕАЛИЗАЦИЯ ETL-ТЕХНОЛОГИИ ДЛЯ ОЗЕРА ДАННЫХ ЭТНОКУЛЬТУРНОГО БАНКА КОРЕННЫХ МАЛОЧИСЛЕННЫХ НАРОДОВ СЕВЕРА

В статье предлагается решение для реализации ETL-процесса и озера данных этнокультурного банка коренных малочисленных народов Севера, агрегирующего информационного ресурса по языкам и культурам коренных малочисленных народов Севера, способного стать базой для использования NLP-технологий и автоматизации процесса обработки текстов на этих языках.

Ключевые слова: обработка данных, озеро данных, ETL, NLP, этнокультурный банк.

V. V. Burlutskiy, V. A. Baluev, M. I. Izert, M. A. Rusanov

IMPLEMENTATION OF ETL-TECHNOLOGY FOR THE DATA LAKE OF THE ETHNOCULTURAL BANK OF THE INDIGENOUS MINORITIES OF THE NORTH

The article proposes a solution for the implementation of the ETL process and the data lake of the Ethnocultural Bank of the Indigenous peoples of the North. The Ethnocultural Bank is an aggregating information resource on the languages and cultures of indigenous peoples of the North, which can become the basis for using NLP technologies and automating the processing of texts in these languages.

Key words: data processing, data lake, ETL, NLP, Ethnocultural Bank.

Введение

В Российской Федерации, по данным Института языкознания РАН [1], для общения используется 295 языков и диалектов, некоторые из них находятся на грани исчезновения. Сохранение языка является ключевым условием сохранения культуры народа, однако в период глобализации оригинальные языки малых народов оказались на грани исчезновения. ЮНЕСКО объявило 2019 год Международным годом языков коренных народов для привлечения внимания к возможной утрате родных языков и необходимости их поддержки и активизации, а в 2010 году этой организацией был выпущен «Атлас языков мира, находящихся под угрозой исчезновения», в котором указывалось, что около 2500 языков во всем мире находится под угрозой исчезновения.

Важнейшее значение в решении проблемы сохранения языков и культур коренных малочисленных народов в наше время приобретают современные информационные технологии. Они представляют возможности не только хранения и предоставления доступа к огромным массивам данных по лингвистике, этнографии, культурологии, но и возможности использования методов искусственного интеллекта и машинного обучения для решения задач автоматической обработки естественных языков и информационного поиска.

Так, в начале 2020 года «Яндекс» включил якутский язык в сервис «Яндекс. Переводчик» [2], который позволяет автоматически довольно точно перевести фразу этого языка на любой другой язык, представленный в переводчике, и наоборот. Помимо якутского языка в «Яндекс. Переводчик» включены еще несколько языков народов России: марийский, горномарийский, удмуртский, чувашский удмуртский и др.

Для применения методов автоматической обработки естественного языка (далее NLP, Natural Language Processing) требуется корпус размеченных текстов, который представляет данный язык на определенном этапе его существования и во всём многообразии жанров, стилей, территориальных и социальных вариантов. Основная проблема, с которой сталкиваются исследователи при использовании методов NLP для обработки малоресурсных языков коренных народов [3; 4], – это недостаточное количество текстов для создания полноценного корпуса, а также их «однобокость», т. е. перекос количества текстов в сторону одного жанра или одного диалекта. Поэтому первоочередной задачей применимости NLP является создание информационного ресурса и комплекса программных инструментов для сбора, унификации и разметки сбалансированного репрезентативного набора текстов на данном языке.

Идея создания электронного банка для доступа к культурному наследию коренных финно-угорских народов, интегрирующего специализированные базы и хранилища о языке и культуре финно-угорских народов России и мира в целом с перспективой применения методов автоматической обработки естественного языка, легла в основу проекта этнокультурного банка финно-угорских народов, инициированного губернатором Ханты-Мансийского автономного округа – Югры на международном форуме «Год языков коренных народов в России».

Кроме этнокультурного банка, технологические решения по хранению данных будут описаны в данной статье ниже, отметим еще ряд успешных отечественных проектов по сохранению языков малочисленных финно-угорских народов. Во-первых, это ресурс “Финно-угория” [5] Финно-угорского культурного центра РФ. Данный ресурс имеет обширный набор данных по различным финно-угорским народам. Однако данные на ресурс вносятся в ручном режиме, что означает их низкую скорость актуализации, а также небольшой охват источников. Во-вторых, информационные ресурсы Межрегиональной лаборатории информационной поддержки функционирования финно-угорских языков [6]. Коллектив лаборатории собрал корпус языка коми, насчитывающий более 60 миллионов словоупотреблений, а также разработал множество инструментов и сервисов (раскладки клавиатур, несколько электронных словарей, модули проверки орфографии и др.), позволяющих интегрировать языки коренных малочисленных народов в современную информационную среду.

Интерес с точки зрения использования для решения задач NLP для языков финно-угорских народов представляют также ресурсы порталов «Культура и язык народов коми», «Культура и язык народов Удмуртии», «Культура и язык народов Республики Марий Эл».

Постановка задачи

Создание агрегационных информационных систем, интегрирующих тексты малоресурсных языков финно-угорских народов и ресурсы об их культуре, имеет ряд ключевых особенностей. А именно: малый объем текстов, в основном фольклорной тематики; неустоявшаяся орфография и лексика, в том числе из-за изменения алфавитов этих языков; небольшое число носителей языка, из которых только часть владеет письменным языком; фрагментарность доступных информационных ресурсов, информация в которых слабо структурирована и зачастую имеет вид скан-копий, которые не являются машиночитаемыми.

Создание этнокультурного банка, агрегирующего информационного ресурса по языкам и культуре коренных малочисленных народов Севера, способного стать базой для использования NLP-технологий и автоматизации процесса обработки текстов на этих языках, требует решения следующих задач:

- разработка технологии хранения для слабоструктурированных специализированных данных из разнородных источников, поддерживающей как минимум эффективный полнотекстовый поиск;
- проектирование ETL-процесса для загружаемых данных и разработка технологического решения для его реализации, включая решения по парсингу и унификации данных;

– разработка комплекса программных инструментов для сбора, унификации и разметки текстов на различных языках финно-угорской группы;
– разработка web-портала, позволяющего использовать реализованный функционал по хранению и обработке текстовых и иных ресурсов этнокультурного банка.

Описание технологии хранения данных этнокультурного банка

В качестве базового решения для хранения данных этнокультурного банка была выбрана технология озера данных [7]. Озеро данных – это метод хранения неструктурированных разнородных данных в их исходном формате. Идея озера данных заключается в том, чтобы иметь единое логически определенное хранилище всех данных, включая как необработанные исходные данные, так и предварительно обработанные и частично структурированные.

Выбор данной структуры обусловлен тем, что в системе предполагается ввод данных из самых различных источников, для которых невозможно заранее описать тип данных, его формат и свойства. Использование стандартных методов хранения данных, например, реляционной базы данных, приведет к необходимости постоянной модернизации архитектуры и структуры хранилища в случае появления дополнительных свойств или характеристик данных, что приведет к серьезным проблемам при развитии и функционировании системы.

Слабой стороной озера данных является сложность представления данных. Так, если необходимо сформировать представление данных о каком-либо человеке, то нужно убедиться, что в озере данных присутствуют минимально необходимые данные: имя, фамилия, национальность (язык). Однако озеро данных не может гарантировать целостность данных, а написание инструментария валидации весьма трудоемко, поэтому было принято решение использовать стандартные средства реляционной базы данных для решения данной задачи. Таким образом, в системе организуется хранение структурированных и неструктурированных данных.

Неструктурированные данные располагаются в таблицах базы данных data, data_type, metadata, metadata_type с учетом таблиц их связей (см. рисунок 1).

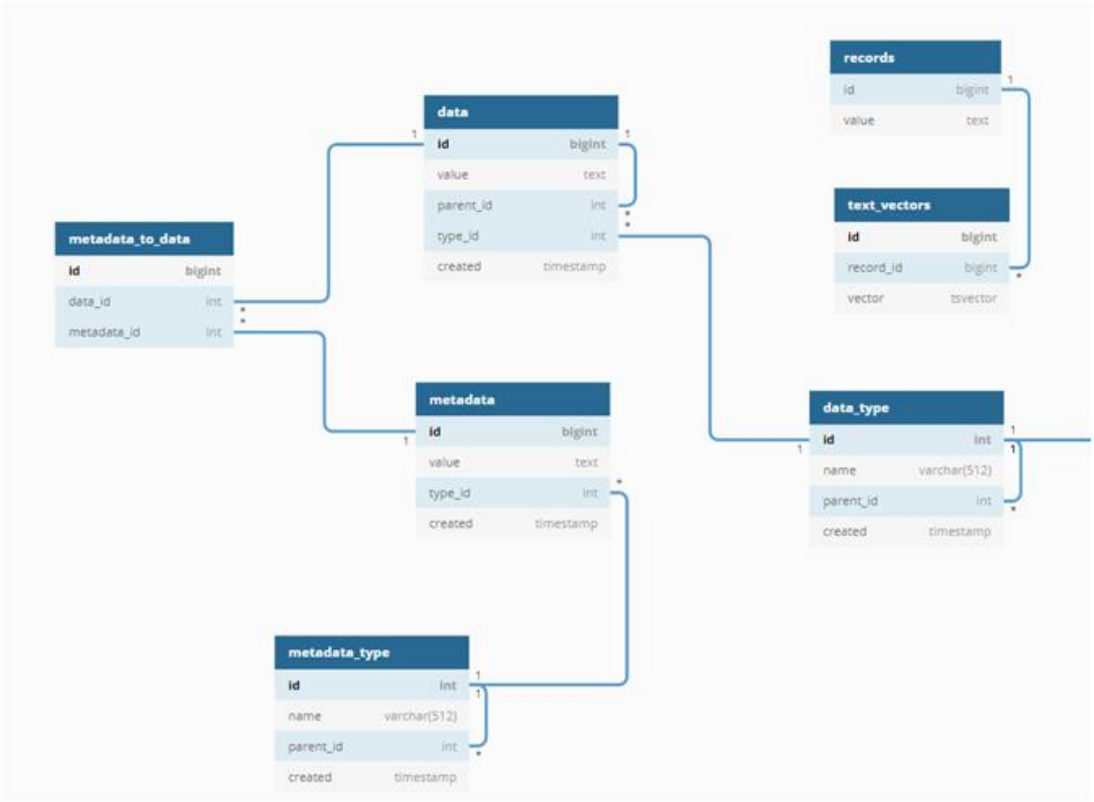

Рисунок 1 – Реализация хранения неструктурированных данных

Таблица data содержит информацию о самом элементе, загружаемом в озеро данных. Запись в таблице данных должна указывать на тип записи из таблицы data_type. Например, если вносится запись о каком-либо человеке, то тип данных должен соответствовать записи из таблицы data_type «Персоналия», а поле value должно содержать ФИО человека. Если же нужный тип отсутствует в таблице data_type, то предварительно его нужно добавить.

Иерархия реализована с помощью указания родительского элемента. Корневые элементы родителя не имеют. Одноранговая иерархия реализована с помощью таблицы link, с элементами которой связаны элементы из data. Одноранговая связь позволяет организовать хранение версий для переводов одного материала на различные языки.

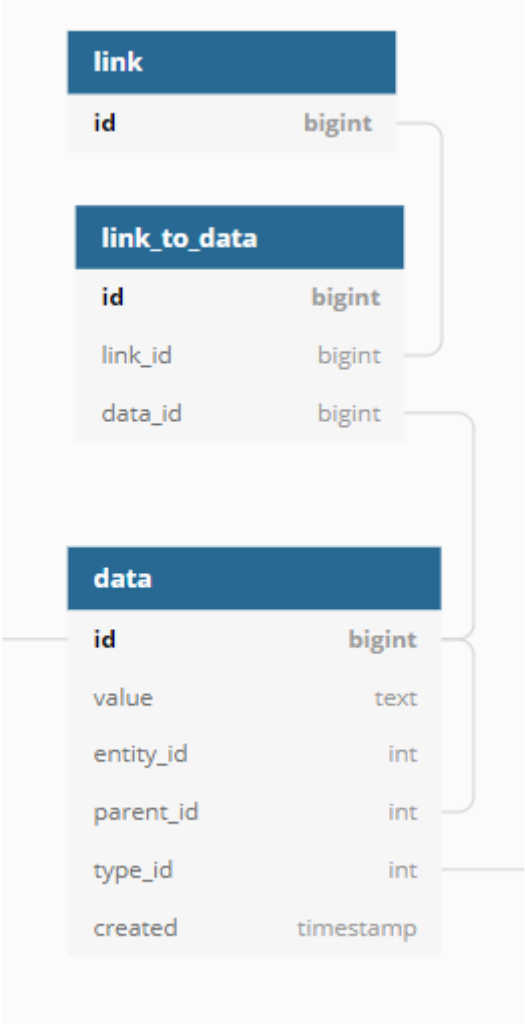

Рисунок 2 – Реализация одноранговой иерархии

Следующим шагом после внесения записи в таблицу data является добавление метаданных, которые привязываются к записи в таблице data. Метаданные хранятся в таблице metadata и содержат в себе информацию о типе метаданных, который представлен в таблице metadata_type, а также содержание самого описания метаданных в поле value.

Таблицы data и metadata наследуются от таблицы record. В таблице record имеется поле value, которое и наследуется дочерними таблицами. Это реализовано с целью внедрения полнотекстового поиска, а также для добавления индексов в базу данных (что позволит значительно ускорить полнотекстовый поиск).

Пример занесения неструктурированных данных представлен ниже.

В таблицу data помещается основная запись, содержащая основное значение и информацию о родительском элементе, типе записи и времени создания.

Таблица 1

data				
id	**Value**	**created**	**parent_id**	**type_id**
1	Иванов Иван Иванович	10-06-2020	Null	1

Таблица 2

data_type			
id	**name**	**parent_id**	**created**
1	Персоналия	Null	10-05-2020 01:00:00

После в таблицу **metadata** помещается описание элемента справочника. Для связывания одного объекта с несколькими справочниками используется таблица **metadata_to_data**. В данном примере объект *Иванов Иван Иванович* связан со справочником персоналия.

Таблица 3

metadata			
id	**Value**	**created**	**type_id**
2	Иванов Иван Иванович	10-06-2020 01:00:00	100
3	01.01.1951	10-06-2020 01:00:00	101
4	Ненецкий	10-06-2020 01:00:00	108
5	говорит и пишет	10-06-2020 01:00:00	102
6	с. Варьёган	10-06-2020 01:00:00	103
7	с. Варьёган	10-06-2020 01:00:00	104
8	Ненецкий	10-06-2020 01:00:00	105
9	Родился в с. Варьёган, закончил 6 классов Варьеганской школы	10-06-2020 01:00:00	106

metada_type		
id	**name**	**created**
100	ФИО	10-06-2020 01:00:00
101	дата рождения	10-06-2020 01:00:00
102	уровень владения	10-06-2020 01:00:00
103	место рождения	10-06-2020 01:00:00
104	место жительства	10-06-2020 01:00:00
105	принадлежность к коренным народам	10-06-2020 01:00:00
106	биография	10-06-2020 01:00:00

Хранение структурированных данных устроено по стандартным принципам реляционной базы данных. На данный момент осуществляется хранение следующих сущностей и их связей: народы; файлы; этнические игры; языки; СМИ. Схема реляционной базы данных представлена на рисунке 3.

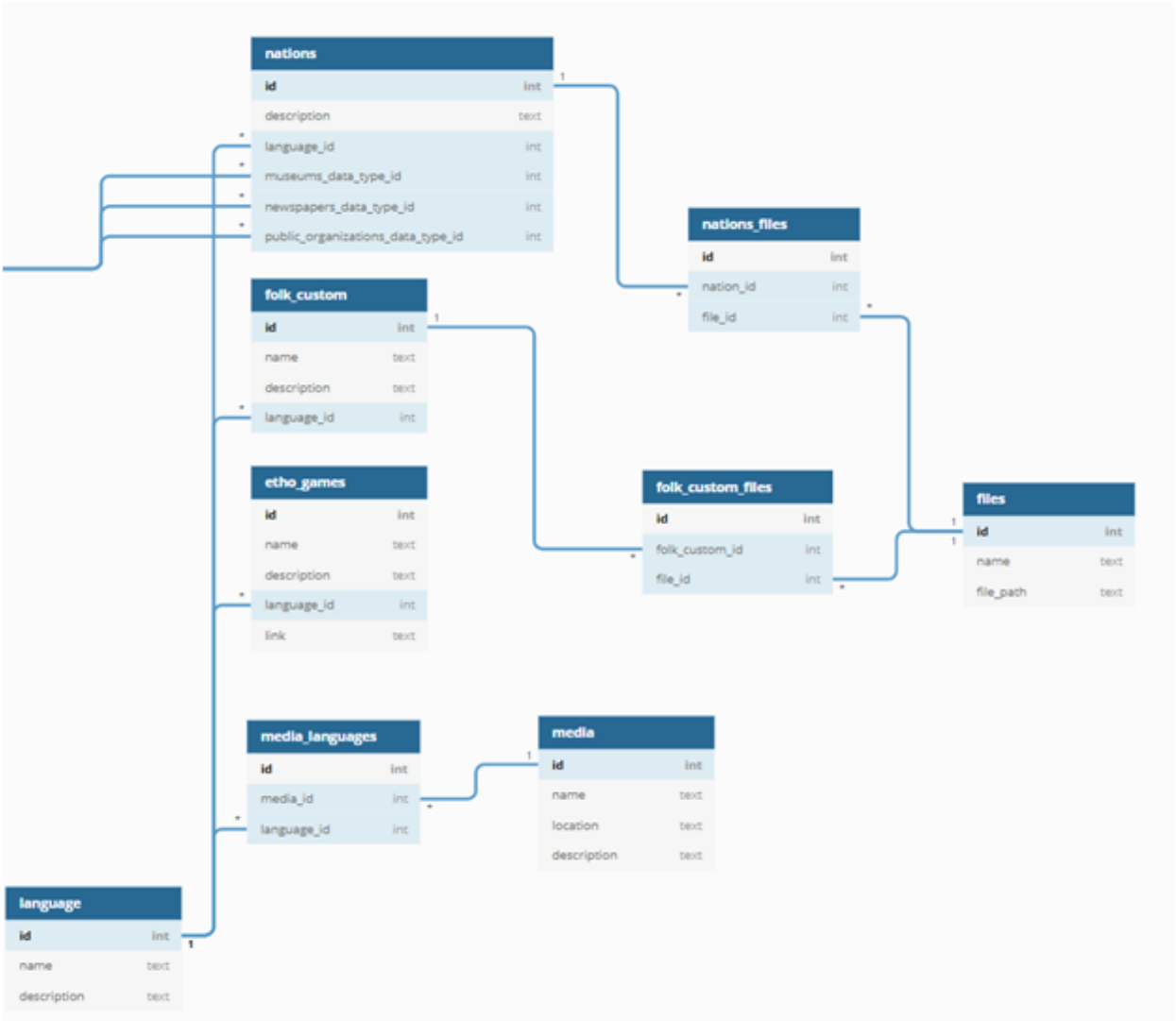

Рисунок 3 – Реализация хранения структурированных данных

Описание процесса ETL этнокультурного банка

Процесс ETL (Extract, Transform, Load) [8] представлен на рисунке 4 и начинается с загрузчиков данных. Ряд автоматизированных модулей осуществляет сбор данных из информационных ресурсов этнокультурного направления. На данный момент осуществляется сбор следующих классов данных: газеты, ресурсы библиотек, персоналии и записи депозитария. Полученные данные передаются на обработку системе в виде json-объектов.

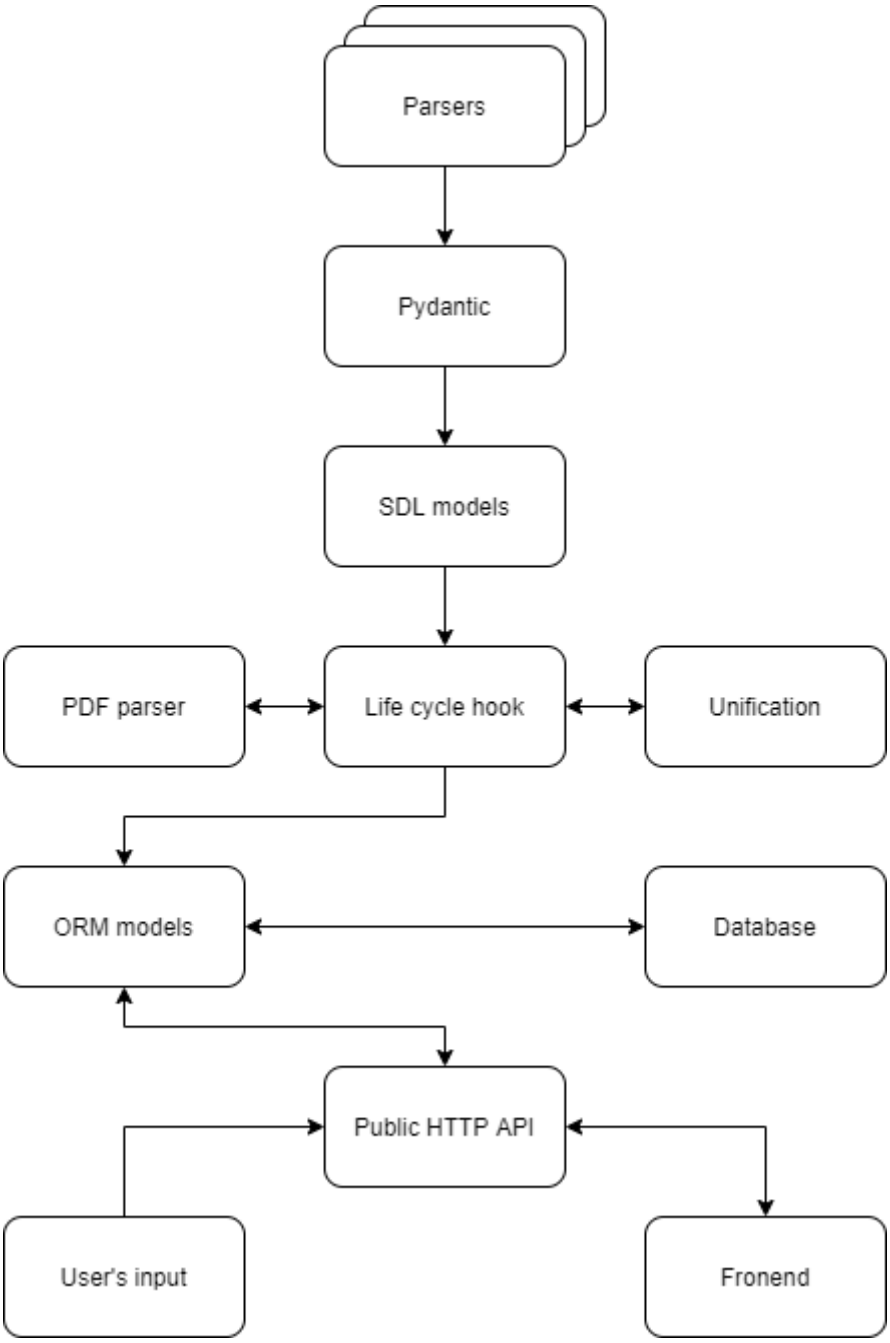

Рисунок 4 – Схема процесса ETL

Полученные данные проходят валидацию с помощью Pydantic с целью исключения ошибок формирования записи. После чего данные переводятся в формат базы данных, сохраняя структуру и связь элементов. С помощью life cycle hook определяется завершение предыдущего этапа и происходит обработка входных данных, а именно унификация и извлечение текста из файлов формата PDF. Под унификацией понимается приведение написания текстов различных вариантов языка к единому для возможности поиска и структурирования данных. При унификации создается одноуровневая связь между исходным документом и унифицированным, которая описана ранее.

После унификации данные приводятся к представлению ORM (Object-Relational Mapping) [9]. ORM представления используются для хранения и обработки данных, полученных от API, с помощью которого происходит взаимодействие frontend'а и системы. Добавление пользовательских данных происходит также через работу с ORM, поскольку пользовательский ввод имеет более строгую валидацию и дополнительные проверки на стороне

веб-интерфейса. Просмотр, поиск и внесение пользовательских данных осуществляются с помощью веб-интерфейса, являющегося SPA, разработанным с использованием компонентного подхода и фреймворка Vue.

Заключение

На основе описанных выше технологий был разработан базовый прототип информационной системы «Этнокультурный банк финно-угорских народов» [10], выполняющий роль агрегатора информации о языковом и культурном наследии коренных народов Севера, в частности ханты и манси. Информационная система реализует следующий функционал:

- хранение данных в соответствии с описанной технологией озера данных;
- ETL-процесс, включая парсинг данных из различных источников, и унификацию данных, получаемых от парсеров и загружаемых непосредственно через портал этнокультурного банка данных;
- API для запроса данных из этнокультурного банка;
- web-сервис для работы с этнокультурным банком.

На момент выхода статьи информационная система «Этнокультурный банк финно-угорских народов» прошла все регламентные этапы тестирования и валидации. Главная страница портала этнокультурного банка представлена на рисунке 5.

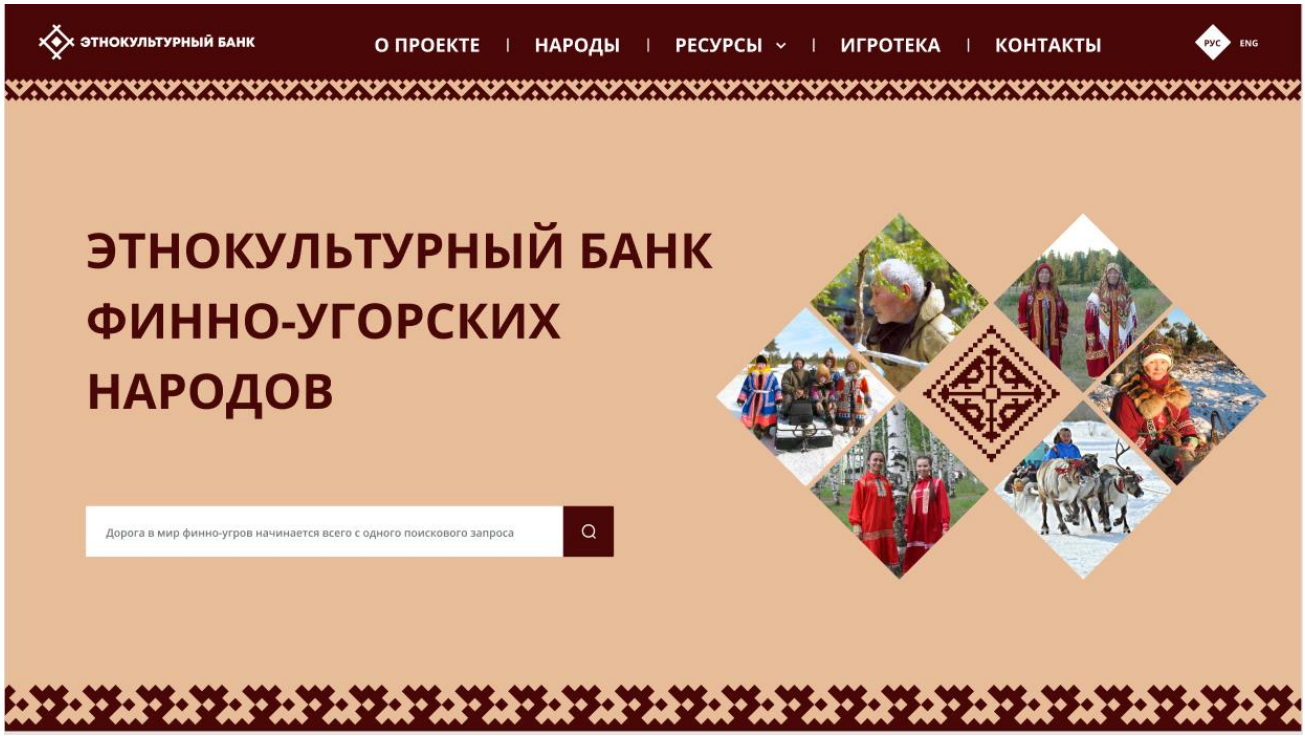

Рисунок 5 – Главная страница портала этнокультурного банка

Литература

1. Языки России / Институт языкознания РАН. – URL: https://iling-ran.ru/web/ru/jazykirf (дата обращения: 02.11.2020). – Текст : электронный.

2. Яндекс переводчик. – URL: https://translate.yandex.ru/ (дата обращения: 02.11.2020). – Текст : электронный.

3. Konovalov, V. P. Learning word embeddings for low resource languages: the case of buryat / V. P. Konovalov, Z. B. Tumunbayarova // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. – 2018. – С. 331–341.

4. About Naki. List of research and engineering of NLP for American Native/Indigenous

Languages. – URL: https://pywirrarika.github.io/naki/ (дата обращения: 02.11.2020). – Текст : электронный.

5. Финно-угорский центр Российской Федерации. – URL: http://www.finnougoria.ru/ (дата обращения: 02.11.2020). – Текст : электронный.

6. Межрегиональная лаборатория информационной поддержки функционирования финно-угорских языков. – URL: https://fu-lab.ru/laboratoriya (дата обращения: 02.11.2020). – Текст : электронный.

7. Майер-Шенбергер, В. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим / В. Майер-Шенбергер, К. Кукьер. – Москва : Манн, Иванов и Фербер, 2014. – 240 с. – Текст : непосредственный.

8. Loshin, D. ETL (Extract, Transform, Load) / D. Loshin // Business Intelligence. – 2nd ed. – Morgan Kaufmann, 2012. – 400 p.

9. Flex 4. Рецепты программирования / Дж. Ноубл, Т. Андерсон, Г. Брэйтуэйт [и др.]. – Санкт-Пепербург : БХВ Петербург, 2011. – 706 с. – Текст : непосредственный.

10. Этнокультурный банк финно-угорских народов. – URL: ethno.uriit.ru (дата обращения: 02.11.2020). – Текст : электронный.