Министерство науки и высшего образования РФ
Федеральное государственное автономное
образовательное учреждение высшего образования
«**СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ**»

Гуманитарный институт
Кафедра информационных технологий
в креативных и культурных индустриях

УТВЕРЖДАЮ
И. о. заведующего кафедрой

_____ М. А. Лаптева

«_____» _____ 2023 г.

БАКАЛАВРСКАЯ РАБОТА

Система автоматизированного получения и анализа открытых данных культурного наследия.

Направление подготовки: 09.03.03 Прикладная информатика

Наименование программы: 09.03.03.30 Прикладная информатика

Руководитель проф., д-р тех. наук О. А. Антамошкин

Выпускник К. Д. Кожин

Консультанты

Нормоконтролер Нигматуллин И.Р.

Красноярск 2023

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В настоящее время культурные объекты становятся все более доступными. Агрегаторы цифрового культурного наследия по всему миру предоставляют новые возможности для исследователей и обычных пользователей. Яркими представителями агрегаторов культурного наследия являются такие сайты как Europeana и Метрополитен-музей. Эти ресурсы предоставляют обычным пользователям возможности для поиска и изучения культур разных народов, а ученым и исследователям технологии API (Application Programming Interface), позволяющие работать с большими объемами культурных данных и создавать собственные наборы данных для исследований.

В нашей стране существует Государственный каталог Музейного фонда РФ, который является крупнейшим агрегатором культурного наследия России. На данный момент он включается в себя коллекцию, состоящую из 37 миллионов объектов. На сайте Госкаталога имеются возможности просмотра и поиска объектов культурного наследия по различным параметрам. Однако, существует ли доступный и автоматизированный способ получения набора данных объектов культурного наследия России для исследователей? В данной работе мы утверждаем, что использование данных из Госкаталога для проведения исследований сейчас не представляется возможным без создания программного кода, что существенно снижает процент исследований о культурном наследии России. Поэтому данная работа посвящена созданию программы для автоматизации процесса получения метаданных культурных объектов Госкаталога, а также их обработки и визуализации.

Актуальность работы обусловлена труднодоступностью и сложностью работы с метаданными культурных объектов Госкаталога для исследователей.

Объект исследования – цифровая коллекция объектов культурного наследия Государственного каталога музейного фонда Российской Федерации.

Предметом исследования являются технологии сбора, обработки и визуализации метаданных объектов культурного наследия Государственного каталога музейного фонда России

Цель настоящей работы – создать программу для автоматизированного сбора, обработки и визуализации культурных данных, а также показать примеры ее использования для исследований.

Задачи исследования:

– Изучить существующие подходы к организации данных в цифровых коллекциях;

– Провести анализ и оценку существующих методов получения данных из коллекции Государственного каталога Музейного фонда РФ;

– Разработать программное решение для автоматизированного получения, обработки и визуализации метаданных культурных объектов;

– Продемонстрировать на реальных примерах пользу созданного нами программного продукта.

Выпускная квалификационная работа состоит из введения, трёх глав, заключения, списка использованной литературы и приложения.

В первой главе, состоящей из четырех параграфов, были рассмотрены цифровые музеи и цифровые коллекции на примере двух крупных агрегаторов культурного наследия: Европеаны и Метрополитен музея. Была поднята проблема работы с большими объемами данных, а также были описаны преимущества решения этой проблемы с помощью применения технологии API. Далее были рассмотрены все возможные способы взаимодействия с культурными данными коллекции Государственного каталога Музейного фонда РФ. Изучив все недостатки этих способов взаимодействия с данными, было принято решение о создании собственного программного решения.

Во второй главе, состоящей из четырех параграфов, был определен необходимый набор функций в создаваемой программе. Далее были описаны все разработанные модули программы SGAT: модуль сбора данных, модуль предобработки данных, модуль визуализации данных. Кроме того, каждый модуль был протестирован на наборе данных Красноярского краевого краеведческого музея по запросу «Красноярск».

В третьей главе, состоящей из двух параграфов, были продемонстрированы два примера реального применения программы SGAT в разных областях. В первом примере рассматривается задача создания набора данных для обучения нейросети. Во втором примере рассматривается изучение пропусков и распределения культурных данных коллекции Госкаталога.

Заключение кратко излагает результаты исследования. В список литературы включены библиографические данные об источниках, использованных в работе.

1 Проблема работы с большими объемами культурных данных

1.1 Цифровые музеи и цифровые коллекции

На протяжении многих лет исследователи всего мира стремятся сохранить культурную и научную историю в музеях, библиотеках и архивах. Однако в настоящее время существует новый способ сохранения и распространения нашего наследия — это создание цифровых коллекций и цифровых музеев. Цифровые музеи — это виртуальные музеи, которые разрабатываются с целью сохранения информации о культурном наследии. Цифровые музеи используют методы оцифровки, хранения и показа культурных и научных объектов на веб-платформах, что дает людям возможность легко получать доступ к этим объектам. А также, при утрате физического оригинала объекта, с помощью таких цифровых архивов, мы все равно сможем его изучить и рассмотреть.

Данные в цифровых музеях представлены в виде цифровых коллекций, которые в свою очередь являются наборами электронных записей, включающих в себя изображения, видео, метаданные, 3D-модели, документы и другие данные, связанные с объектами культурного наследия.

Ярким примером цифрового музея можно назвать Европеану [1], которая является крупным агрегатором европейской культуры. Сейчас коллекция Европеаны насчитывает более 55 миллионов объектов культурного наследия. На главной странице коллекции (Рисунок 1) имеются различные фильтры для удобного поиска необходимых объектов, такие как выбор типа объекта, цвета объекта или музея, где находится объект. Таким образом, Европеана является ценным ресурсом для сохранения и поиска культурных данных, предоставляя широкий доступ к множеству ценных материалов.

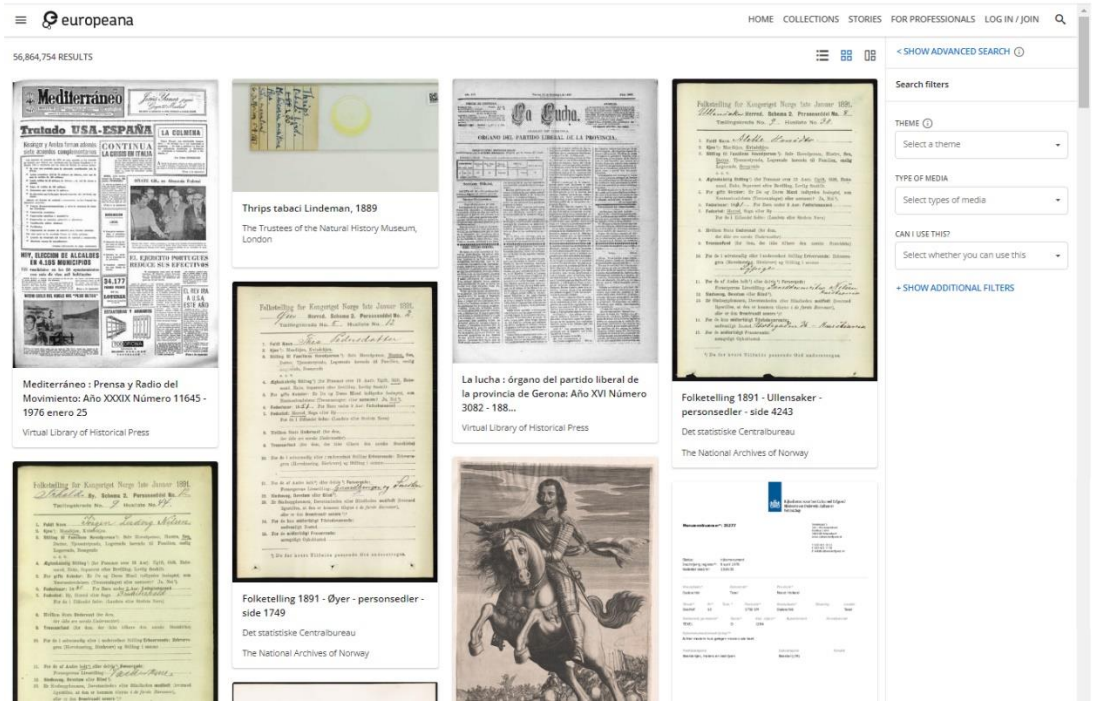

Рисунок 1 – Страница поиска объектов по коллекциям сайта Европеана.

Если говорить о менее крупных цифровых коллекциях, то можно выделить сайт музея Метрополитен [2], на котором в данный момент хранится более 490 тысяч объектов искусства. Метрополитен музей, как и Европеана, имеет функцию фильтрации объектов, но, так как каждый размещенный объект тегируется, поиск необходимых объектов на сервисе значительно облегчен. С помощью параметров можно выполнять поиск, выбирая тип объекта или страну происхождения. Также имеется функция ранжирования объектов в определенном порядке, что позволяет по-разному изучать наполненность сайта (Рисунок 2).

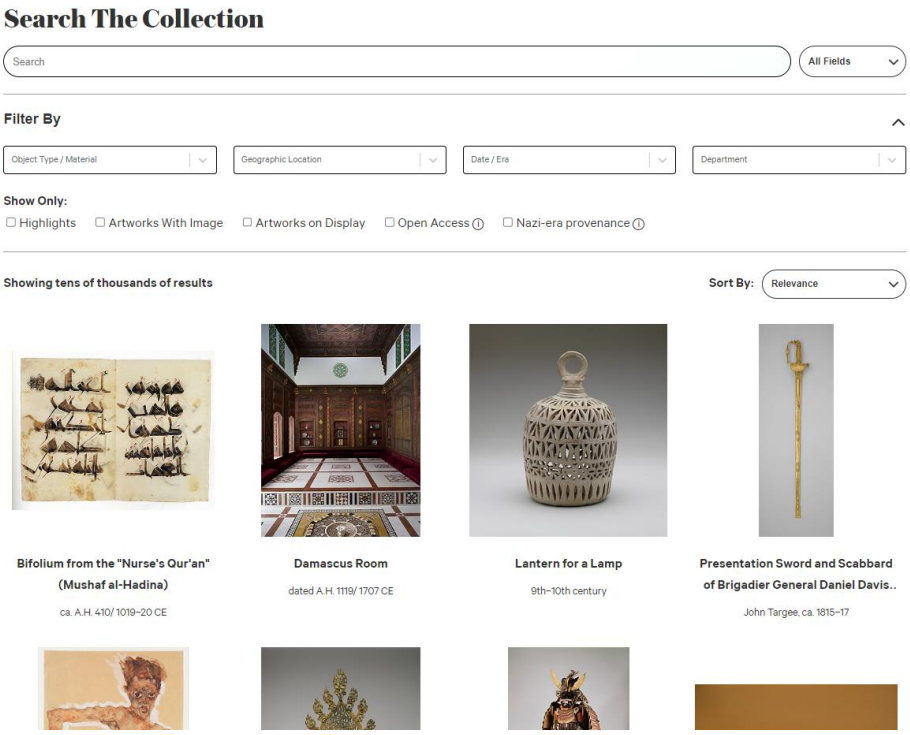

Рисунок 2 - Страница поиска объектов по коллекциям сайта Метрополитен.

Представленные выше цифровые музеи являются хорошими примерами организации культурных данных в веб-среде. С помощью открытых площадок исследователи и простые пользователи могут изучать культурные особенности разных народов, узнавать больше об истории человечества. Именно поэтому важно создавать и пополнять такие коллекции, а также хранить данные так, чтобы их можно было удобно исследовать.

1.2 Использование API для взаимодействия с большими объемами культурных данных

В какой-то момент развития цифровых музеев коллекции в них стали исчисляться миллионами объектов. Стало необходимо найти способ взаимодействия с таким большим массивом данных. На помощь цифровым музеям пришел такой способ взаимодействия с данными как API. Application Programming Interface — это интерфейс, который позволяет программистам взаимодействовать с функциональностью какого-либо программного обеспечения или сервиса. В контексте цифровых коллекций музеев,

использование API упрощает доступ к огромным объемам данных, а также облегчает поиск и анализ информации. Благодаря этому, исследователи и любители искусства могут более эффективно работать с цифровыми коллекциями, что позволяет расширять наши знания о культурном наследии.

Если говорить о научной пользе применения API в цифровых музеях, то можно найти много примеров использования такого метода взаимодействия с культурными данными для проведения исследований. Так, для сбора данных в работах [3,4,5,6] авторы используют API сайта Европианы, а в работах [7,8,9] авторы используют API сайта Метрополитен музея. Таким образом, мы видим, что этот инструмент востребован в научной среде и позволяет ученым делать исследования на основе культурных данных удобнее и быстрее.

1.3 Агрегатор культурного наследия России

Если говорить о культурном наследии России, то в нашей стране существует такой агрегатор культурного наследия, как Госкаталог [10]. Государственный каталог Музейного фонда России - крупнейший агрегатор культурного наследия России, объединяющий все государственные музеи РФ, содержащий основные сведения о каждом музейном предмете и каждой музейной коллекции, включенных в состав Музейного фонда РФ. На сайте агрегатора публикуются объекты более 3300 музеев России. Коллекция Госкаталога на 2023 год насчитывает более 37 миллионов объектов культурного наследия и продолжает активно пополняться новыми предметами. Также каждый культурный объект отнесен к одному из 15 классификационных разделов (Рисунок 3). На сайте также имеется базовая фильтрация по автору, музею, типу объекта.

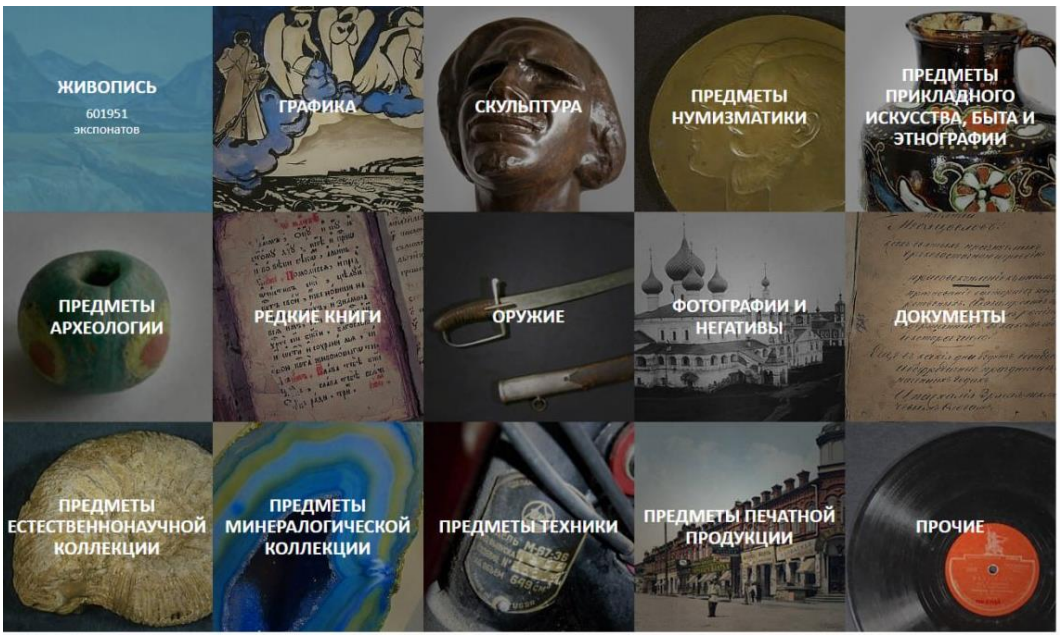

Рисунок 3 - Классификационные разделы коллекции Госкаталога.

Анализируя поисковую выдачу в поисковой системе по научным публикациям Google Scholar [11], мы определили, что на данный момент научных работ, основанных на данных Госкаталога, на порядок меньше, чем работ с использованием данных Европеаны или музея Метрополитен. Вероятно, одной из причин этого является отсутствие способов получения данных на сайте Госкаталога. Так, например, в работе [12] авторы взяли для исследования культурных данных выборку 10%, так как им пришлось заполнять таблицу для последующей визуализации вручную, что затруднило работу с культурными данными. Тем не менее, Госкаталог является незаменимым инструментом для тех, кто ищет информацию о культурных объектах в России.

1.4 Портал открытых данных Минкультуры России

Портал открытых данных Минкультуры России [13] (Рисунок 4) — это веб-сервис предоставляющий доступ к базам данным, которые связаны с культурой и искусством, таким как реестр анимационных организаций или базе данных репертуаров театров и театральных коллективов. Среди баз

данных портала опубликована и необходимая нам база данных коллекции Госкаталога, обновляемая с периодичностью раз в месяц.

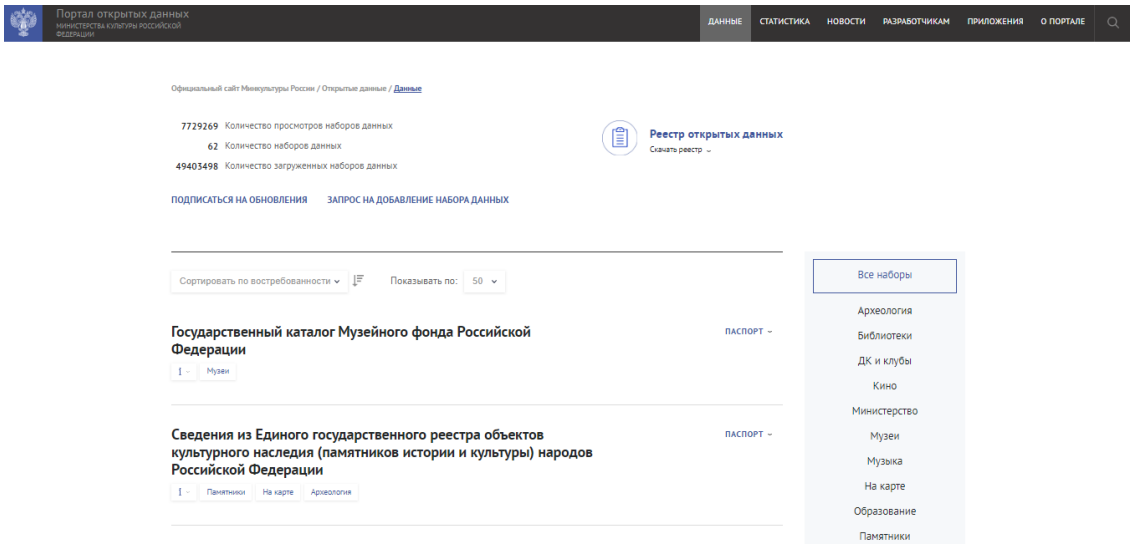

Рисунок 4 – Раздел «данные» портала открытых данных Минкультуры.

Сервис предлагает разработчикам и аналитикам пользоваться программным интерфейсом (API), для этого необходимо зарегистрироваться и получить специальный ключ. Готового программного решения нет, поэтому для извлечения данных необходимо писать программный код. Такой способ подходит не всем, так как для этого нужны специальные технические компетенции.

Также на сайте открытых данных МК РФ размещены три файла, которые содержат в себе метаданные объектов коллекции Госкаталога: JSON.ZIP, JSONS.ZIP, CSV.ZIP. Использование их в исследовательских целях было бы хорошим решением, но при попытке скачать и разархивировать эти файлы мы столкнулись с проблемами. Два архива оказались поврежденными и извлечь из них данные программными средствами не получилось. Оставшийся архив удалось “вылечить” и его можно использовать, например в качестве базы данных для разработки и проверки алгоритмов обработки данных. Однако данный архив содержал данные всего о 20-ти миллионах объектов, что составляло 55% от всего

количества объектов Госкаталога и его нельзя использовать для исследований конкретных исторических и культурных процессов, которые требуют точные параметры поиска объектов, а в данном случае мы имеем неполную базу данных сформированную не по конкретному запросу, а по порядковому номеру объектов.

По запросам на русском и английском языках в поисковой системе по научным публикациям Google Scholar нам не удалось найти работы, где сбор культурных данных осуществлялся бы с помощью API сайта открытых данных Минкультуры РФ. Есть работа [14], где авторы используют данные одного из трех архивов для демонстрации работы методов обработки культурных данных Госкаталога, но они также не применяли API сайта открытых данных Минкультуры для получения данных.

Таким образом, на данный момент нет простого и удобного способа для сбора данных из коллекции Госкаталога для исследователей без специальных технических компетенций.

2 Разработка программного решения для работы с культурными данными Госкаталога

2.1 Идея создания программы

Так как на сайте Госкаталога отсутствует API, а API портала открытых данных Минкультуры имеет ряд ограничений, было создано собственное программное решение на языке программирования Python под название SGAT [15], которое позволяет пользователям без особых технических компетенций удобно работать с культурными данными Госкаталога.

Для эффективной работы с данными необходимо было добавить функции предобработки и визуализации данных. Для удобства пользователей, программа была разделена на 3 модуля: сбор данных, предобработка данных, визуализация данных. Такой подход позволяет быстрее работать с данными, когда пользователю необходимо использовать только один модуль. Далее мы описали работу каждого модуля программы SGAT на примере работы с метаданными объектов Красноярского краевого краеведческого музея (КОПУК 111523) по запросу «Красноярск». Также программный код всех алгоритмов программы можно изучить на странице репозитория GitHub [16].

2.2 Модуль сбора данных программы SGAT

Программа SGAT включает в себя 2 способа сбора данных: с помощью API сайта открытых данных Минкультуры РФ (далее, первый способ) и напрямую с сайта Госкаталога (далее, второй способ). Такое решение из двух способов необходимо, так как у каждого из них есть свои плюсы и минусы.

Первый способ работает с помощью библиотеки Requests [17], которая необходимая для HTTP-запросов к сайту открытых данных министерства культуры, на котором хранится пополняемая база данных объектов Государственного каталога Музейного фонда РФ. Так как этот API имеет ограничения, было необходимо с помощью циклов создать алгоритм, который позволяет получить максимальное количество объектов по запросу.

В окне первого способа сбора данных программы SGAT (Рисунок 5) есть параметры запроса, с помощью которых можно конкретизировать запрос. Поиск происходит по названию объекта, поэтому можно выбрать поиск по лемме или по словоформе объекта. Есть возможность ввести название или идентификационный номер музея, чтобы получить объекты из конкретного учреждения. Каждый объект Госкаталога отнесен к одной из 15 тематик, которую также можно выбрать в качестве параметра для поиска.

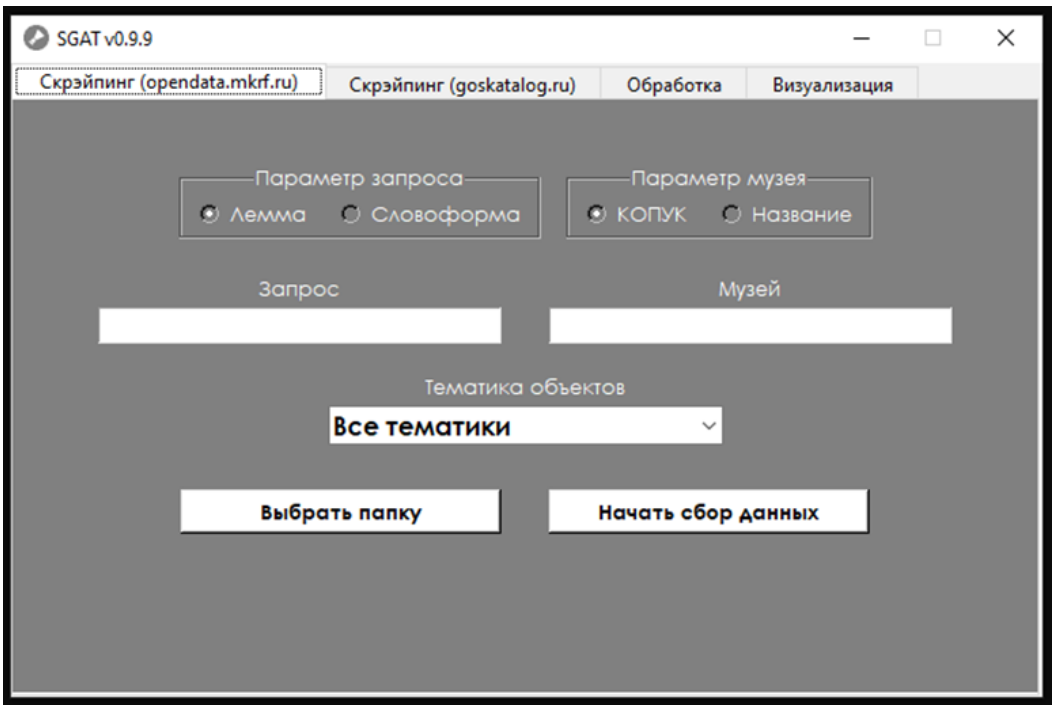

Рисунок 5 - Окно первого способа сбора данных программы SGAT.

После всех тестов алгоритма первого способа сбора данных программы SGAT была составлена Таблица 1, в которой выделены сильные и слабые стороны этого способа.

Таблица 1 - Плюсы и минусы первого способа сбора данных программы SGAT.

Плюсы первого способа сбора данных	**Минусы первого способа сбора данных**
Скорость сбора данных 1000 объектов за 5 секунд 16 полей метаданных в программе (всего с помощью API можно собирать данных по 30+ полям метаданных Параметры запроса (по названию, по музею, по тематике)	За один запрос можно собрать метаданные не более чем 101000 объектов

Таким образом, с помощью первого способа сбора данных можно собрать набор данных с различными настройками поиска, но из-за ограничений по количеству собираемых данных за раз API сайта открытых данных Минкультуры РФ пользователю нужно будет с помощью параметров частями извлекать необходимые данные.

Второй способ сбора данных работает с сайтом Госкаталога напрямую. Так как у него нет собственного API, но данные на нем более свежие и актуальные, необходимо также иметь возможность собирать данные таким способом. Для этого используется открытая библиотека Selenium [18] из-за невозможности использования библиотеки из первого способа для динамически обновляемых сайтов. Сбор данных в нем осуществляется в два этапа. Сначала программе необходимо собрать все ссылки объектов по запросу и сформировать из них список, который записывается в txt файл. Два этапа необходимы, так как пользователь перед вторым этапом может отредактировать по своему усмотрению txt файл со ссылками. С помощью параметров можно конкретизировать запрос, указав количество и ID музея. На втором этапе программа проходит по каждой ссылке и собирает все имеющиеся поля метаданных, в том числе ссылку на объект и ссылку на

изображение объекта при наличии. Окно второго способа сбора данных изображено на Рисунке 6.

Рисунок 6 - Окно второго способа сбора данных программы SGAT.

После всех тестов алгоритма второго способа сбора данных программы SGAT была составлена Таблица 2, в которой выделены сильные и слабые стороны этого способа.

Таблица 2 - Плюсы и минусы второго способа сбора данных программы SGAT.

Плюсы второго способа сбора данных	**Минусы второго способа сбора данных**
Количество объектов для сбора почти не ограничено	Скорость сбора данных 1000 объектов за 25 минут Всего 12 полей метаданных Запрос только по названию объекта

Так, при использовании второго способа сбора данных можно собрать более актуальные данные, но количество собираемых полей метаданных

будет меньше, чем в первом способе, а также сбор данных займет значительно больше времени.

После сбора данных формируется файл в формате CSV с 12 или 16 столбцами в зависимости от выбранного способа сбора данных. Получилось собрать метаданные 16430 культурных объектах. В Таблице 3 можно увидеть пример информации о собранном программой объекте.

Таблица 3 - Пример информации об объекте, собранного первым способом сбора данных.

Название поля метаданных	**Значение поля метаданных**
ID	10034109
Автор	—
Название	Памятная медаль "Красноярскому комсомолу 50 лет". В футляре красного цвета с золотистой застёжкой.
Описание	Круглой формы золотистого цвета.
Период создания	1969 г.
Место создания	РСФСР. Красноярск
Музей	Краевое государственное учреждение культуры "Красноярский краевой краеведческий музей"
КОПУК музея	111523
ID музея	1392
Тематика	предметы нумизматики
Технологии	алюминий, ткань, коленкор, дерево, штамповка, анодирование
Размер	d-45 мм
Начало периода	—
Конец периода	—
Дата регистрации	13.12.2017
Изображения	http://goskatalog.ru/muzfo-imaginator/rest/images/original/6920387

Таким образом, с помощью различных параметров запроса собирается структурированный массив данных, с которым можно работать дальше.

2.3 Модуль предобработки данных программы SGAT

На сайте Государственного каталога Музейного фонда РФ нет единого стандарта заполнения метаданных, поэтому для работы с ними требуется проводить предобработку, чтобы привести всю информацию об объекте в машиночитаемый вид. Для этого в программе SGAT существует модуль предобработки данных (Рисунке 7). Среди параметров в этом модуле имеется выбор предобрабатываемого столбца, а также есть возможность удалить ненужный столбец из загружаемого файла.

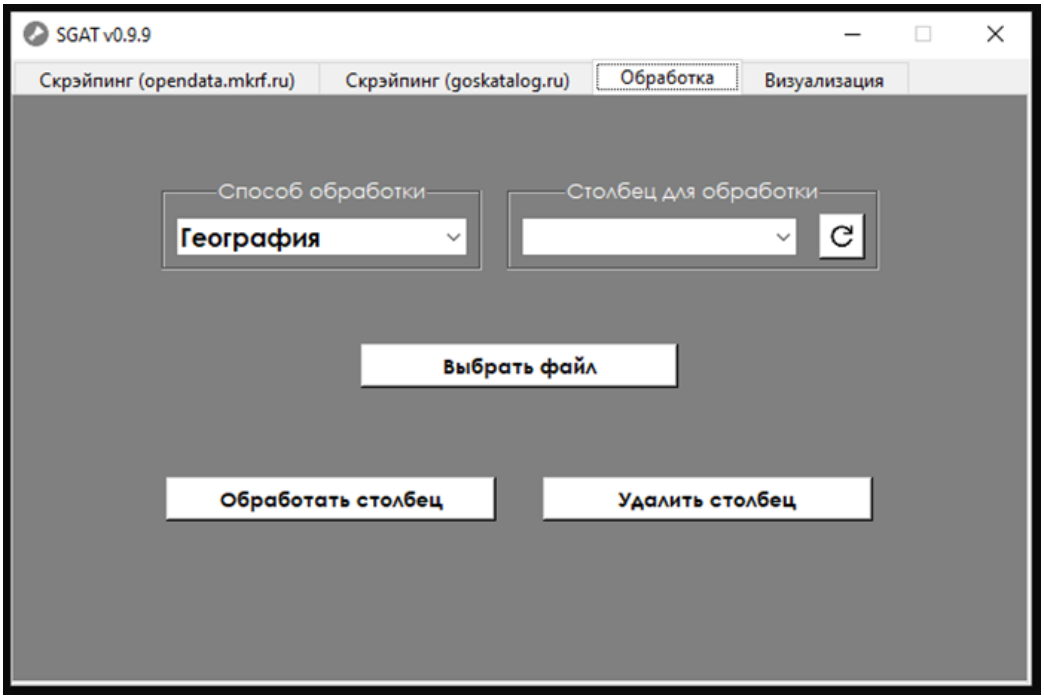

Рисунок 7 - Окно предобработки данных программы SGAT.

Модуль предобработки данных включает в себя четыре типа предобработки данных: предобработка дат, лемматизация слов, поиск коллокаций, предобработка географических названий.

Для визуализации объектов по времени нужно приводить даты к машиночитаемому виду. Если взять за минимальную единицу времени год, то можно выделить 2 типа дат: точную и неточную. Точная дата – это дата, которую без проблем можно представить в виде точки на графике, а неточная дата – это промежуток годов, десятилетий, веков. С помощью алгоритма

далее происходит преобразование двух типов дат. Сначала с помощью библиотеки Pullenti [19] происходит поиск четких дат. После этого необработанные данные проходят через специальный словарь, который можно по желанию пользователя редактировать и дополнять своими правилами преобразования дат (Рисунок 8).

```
1801-1900,xix в.,19,xix век
1901-2000,xx в.,20,xx век,20-й век,xx век,20 в.
2001-2022,xxi в.

1885-1900,конец xix века,кон. xix в.,конец xix в.
1985-2000,конец xx в.,конец xx века
1985-2015,конец 20-начало 21 века,конец xx века начало xxi века
1885-1915,кон. xix - нач. xx вв.,конец xix - начало xx в.,кон. xix в. – нач. xx в.,кон. xix - нач. xx вв

2001-2015,начало xxi века,начало xxi  века

1801-1815,начало xix в.,начало xix века,нач. xix в.,нач. xix века,начало 19 в.,начало 19 века,нач. 19 в.
1901-1915,начало xx в.,нач.20 в.,начало xx века,нач. xx в.,нач. xx века,начало 20 в.,начало 20 века,нач.

1901-1925,перв.четв. xx в.,первая четверть xx в.,первая четверть xx века

1926-1950,втор.четв. xx в.,вторая четверть xx в.

1901-1950,первая половина xx в.,первая половина xx века,1-я половина xx века,перв. пол. xx в.,первая пол

1951-2000,вторая половина xx в.,вторая половина xx века,вторая половина xx века,2-я половина xx века,2 г

1901-1933,перв. треть xx в.

1851-1900,втор. пол. xix в.,вторая половина xix века,вторая половина xix в.,2-я половина xix века


1800-1809,1800-е гг.,1800-е
1810-1819,1810-е гг.,1810-е
1820-1829,1820-е гг.,1820-е
1830-1839,1830-е гг.,1830-е
1840-1849,1840-е гг.,1840-е
```

Рисунок 8 - Фрагмент словаря для преобразования нечетких дат в машиночитаемый формат.

Перевод дат из нечеткого вида в машиночитаемый формат является серьезной проблемой, так как нет определенных правил преобразования дат. Каждый исследователь может преобразовывать такие периоды создания основываясь на своих правилах. Так, для преобразования дат в программе SGAT использовались правила, представленные в Таблице 4.

Таблица 4 – Примеры правил преобразования нечетких дат в машиночитаемый формат.

Исходные данные	**Преобразованные данные**
Начало 20 века	1901-1915
Конец 20 века	1985-2000
Конец 20 века – начало 21 века	1985-2015
Первая четверть 20 века	1901-1925
Вторая четверть 20 века	1926-1950
1940-е гг.	1940
Первая треть 20 века	1901-1933
Первая воловина 20 века	1901-1950
20 век	1901-2000

Таким образом, большую часть дат, записанных человеческим языком, удается определить и представить в виде года или промежутка годов (Таблица 5).

Таблица 5 – Примеры преобразования дат с помощью модуля предобработки данных программы SGAT.

Исходные данные	**Предобработанные данные**
конец 20 в.	1985-2000
втор.четв. 20 в.	1926-1950
1970-е гг.	1970-1979
кон. 19 - нач. 20 вв.	1885-1915
3.04.1958 г.	1958

Далее необходимо разобрать следующий тип преобработки данных программы SGAT. Лемматизация — это процесс приведения слова к его базовой форме (лемме), который позволяет сократить количество различных форм одного слова и улучшить качество анализа текста и поиска по нему. С помощью лемматизации можно получить список самых частотных слова для последующей визуализации. Для этого используется алгоритм, использующий библиотеки Natural Language Toolkit [20] и Pymorphy2 [21]. После этого формируется файл с самыми частотными словами (Рисунок 9), по желанию пользователь может с помощью специального файла обозначить стоп-слова, которые не будут включены в итоговый файл с леммами.

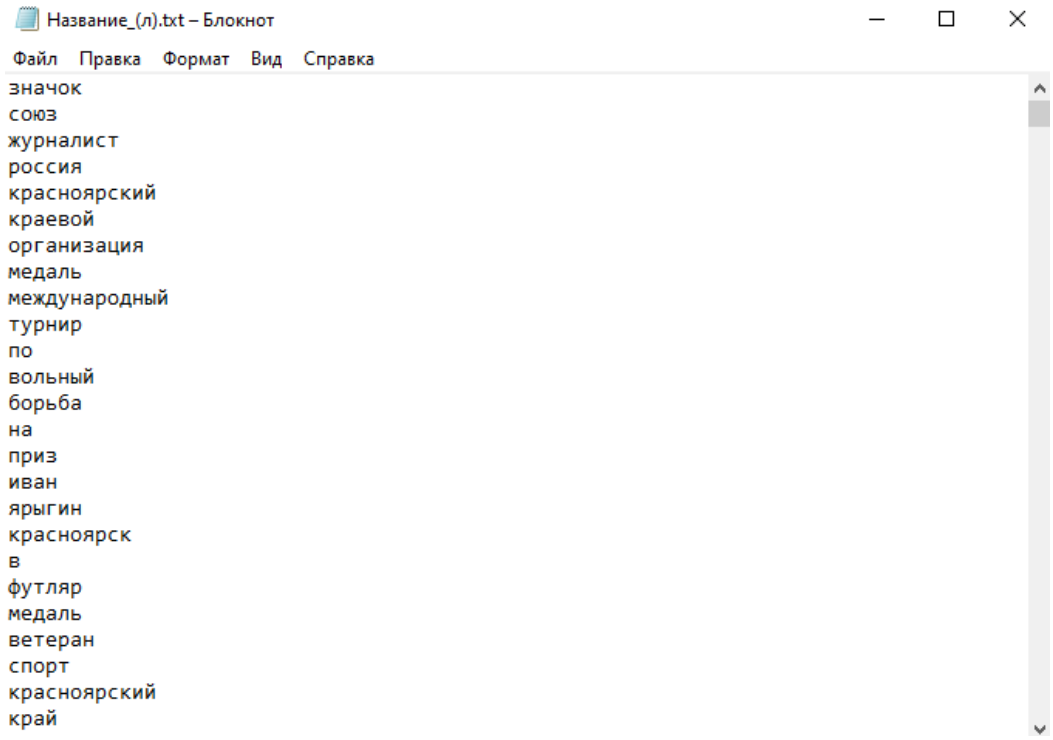

Рисунок 9 – Итоговый txt-файл с лемматизированными словами из названий объектов.

Схожим образом в программе SGAT работает поиск коллокаций. Коллокация — это сочетание двух и более слов, которые часто используются вместе и имеют тенденцию к совместному употреблению в определенном контексте. Для поиска коллокаций был создан алгоритм, использующий те же библиотеки, что и для лемматизации. После завершения работы алгоритма формируется файл со всеми найденными коллокациями (Рисунок 10).

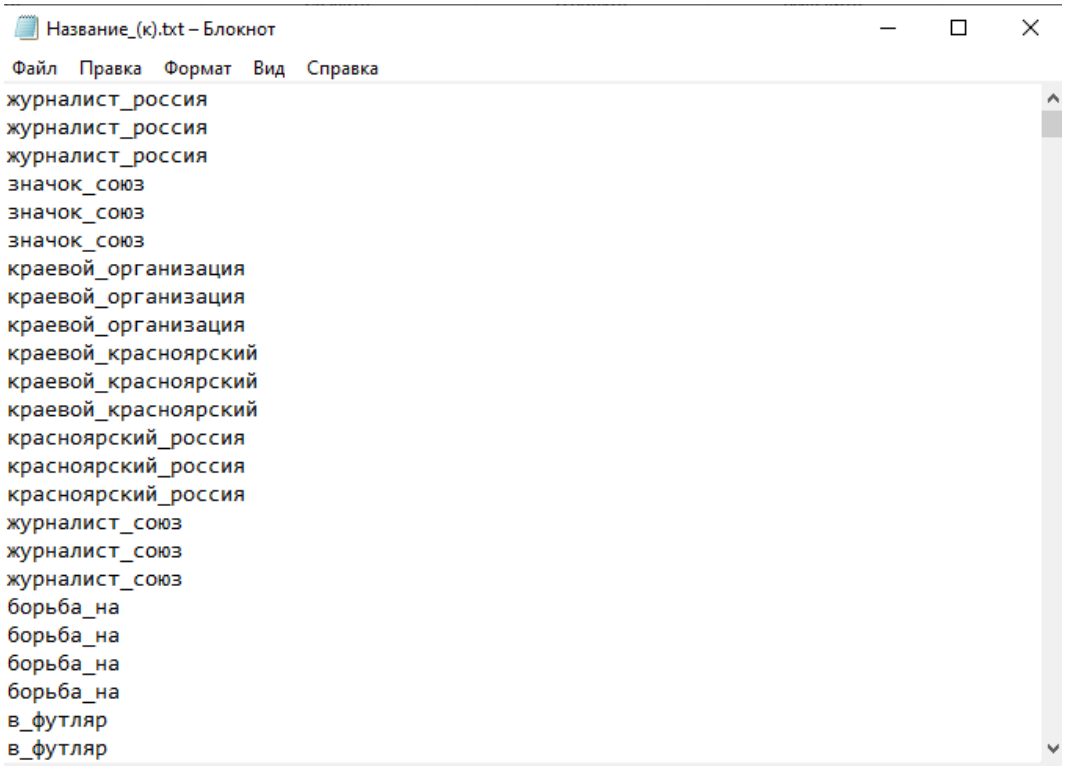

Рисунок 10 - Итоговый txt-файл со всеми найденными коллокациями из названий объектов.

Последним на данный момент типом предобработки данных программы SGAT является предобработка географии. Так как поле «место создания» также не имеет четкий стандартов заполнения, необходимо привести все наименования к единому виду. Алгоритм работает с помощью вышеупомянутой библиотеки Pullenti. Библиотека может распознавать страны, города, деревни, поселки и другие населенные пункты. В Таблице 6 представлены примеры преобразования географических названий.

Таблица 6 – Примеры преобразования географических названий с помощью модуля предобработки данных программы SGAT.

Исходные данные	**Предобработанные данные**
Российская Федерация. Красноярск	РОССИЯ, КРАСНОЯРСК
РСФСР. Красноярск	РСФСР, КРАСНОЯРСК
СССР, г. Красноярск	СССР, КРАСНОЯРСК
Енисейская губ., г. Красноярск	КРАСНОЯРСК
г. Москва	МОСКВА
СССР, Грузинская ССР, Тбилиси	ТБИЛИСИ

Таким образом, с помощью модуля предобработки программы SGAT можно подготовить данные для визуализации.

2.4 Модуль визуализация данных программы SGAT

Визуализация данных важна, так как она помогает ученым из разных областей быстрее и эффективнее понимать большие объемы информации, выявлять закономерности, делать выводы и принимать решения, а также представлять свои результаты более наглядно и доступно для широкой аудитории.

Модуль визуализации программы SGAT (Рисунок 11) включает в себя 4 способа визуализации: гистограмма, облако лемм/коллокаций, график пропусков в данных, график распределения по времени. Визуализация в программе осуществляется за счет библиотеки Matplotlib [22]

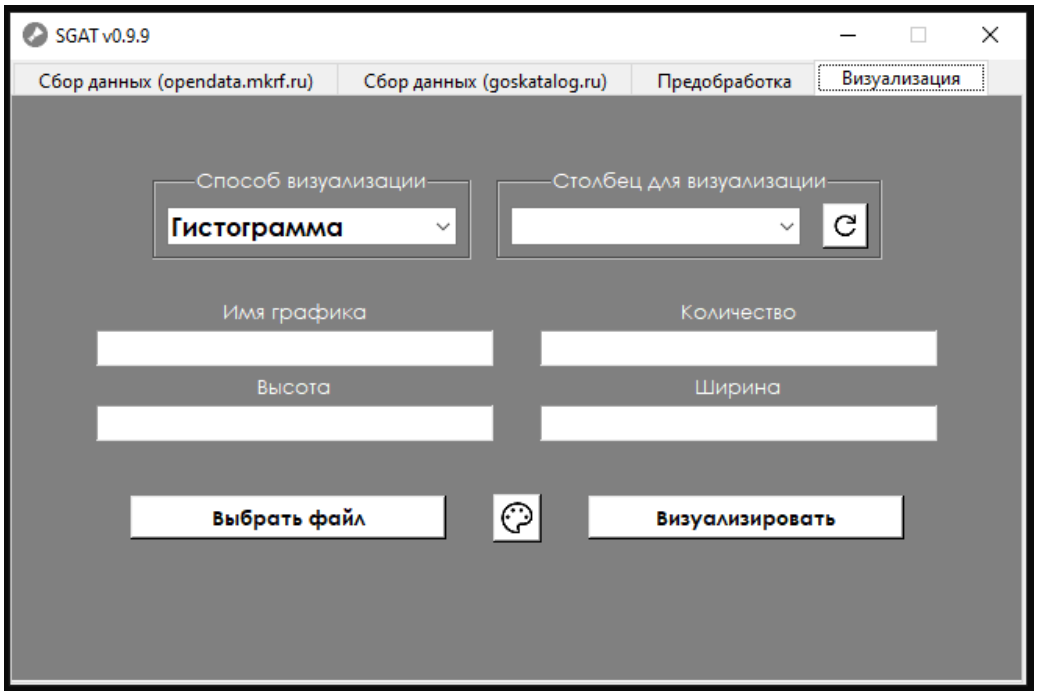

Рисунок 11 - Окно визуализации данных программы SGAT.

Гистограмма — это графическое представление распределения частоты значений переменной. Она может быть полезна для визуализации больших объемов данных и позволяет быстро оценить характеристики распределения, такие как среднее значение, медиана и дисперсия. Применение гистограммы

к культурным данным может быть полезно для составления топов мест создания объектов, тематик объектов и других параметров. Кроме того, гистограмма помогает понять, с какими данными мы имеем дело, что способствует формированию правильных гипотез для дальнейшего исследования. Например, гистограмма может показать, что большинство объектов в выборке являются фотографиями, что может быть важной информацией для исследования. В данном случае (Рисунок 12) с помощью гистограммы мы выяснили, что по запросу Красноярск в Красноярском краевом краеведческом музее большинство объектов являются документами, фотографиями и предметами печатной продукции.

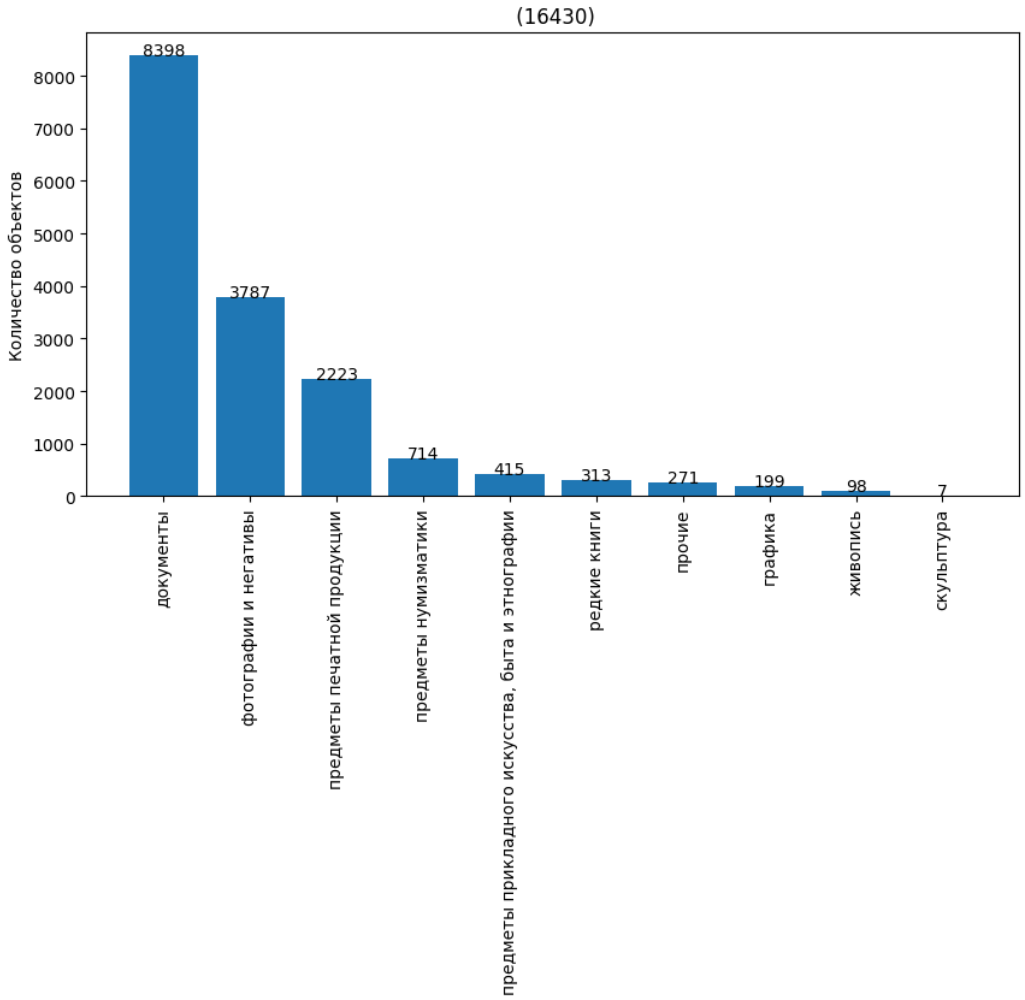

Рисунок 12 – Визуализация в виде гистограммы, показывающая распределение объектов нашей выборки по коллекциям.

Визуализация данных в виде облака слов — это графическое представление текстовой информации, где размер каждого слова пропорционален его частоте в тексте. Такой график может быть использован для анализа частотности слов в различных контекстах, в том числе для выявления наиболее употребляемых слов при заполнении метаданных объектов Госкаталога. Так, с помощью облака слов можно выяснить какие слова преобладают в названиях объектов нашей выборки (Рисунок 13).

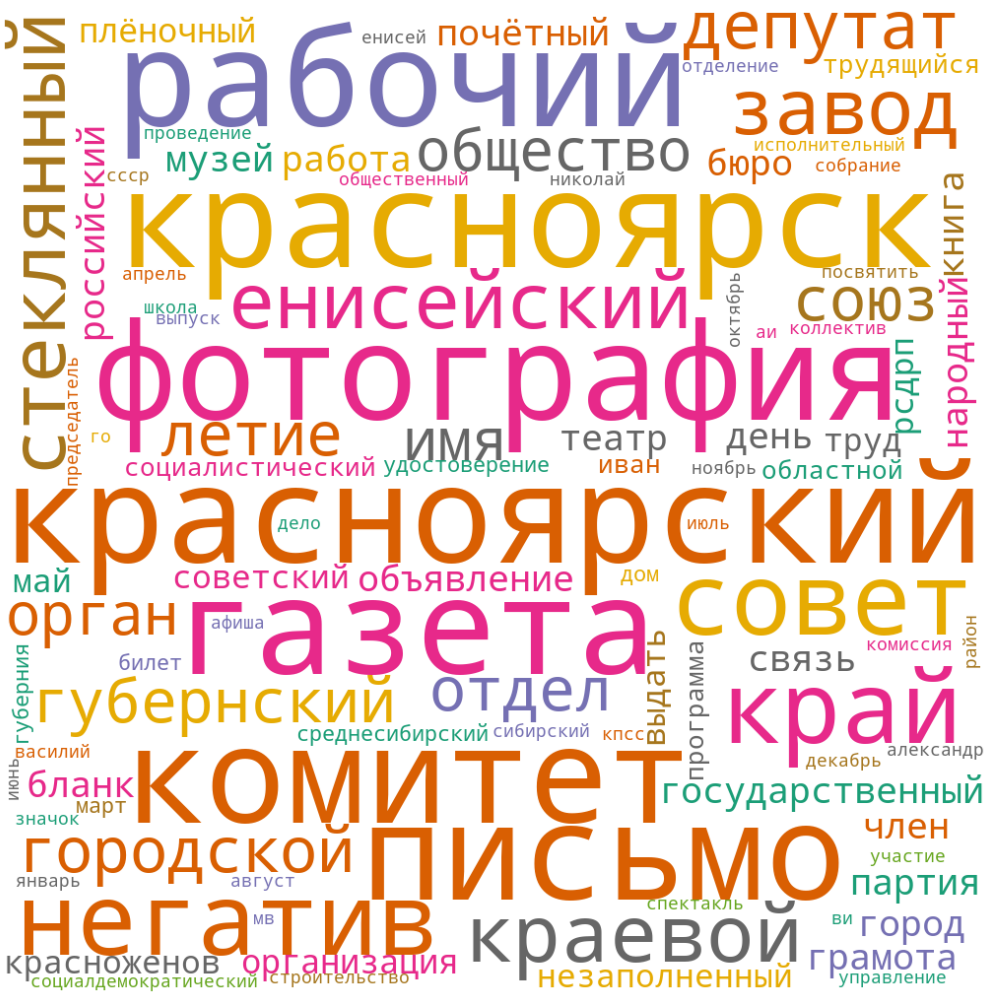

Рисунок 13 - Визуализация в виде облака слов, построенная по названиям объектов нашей выборки.

Анализ словосочетаний позволяет более точно определить контекст, в котором используются ключевые слова. Благодаря этому можно получить более глубокое понимание содержания текста и выделить более

специфические темы, которые обсуждаются в тексте. Таким образом, визуализация коллокаций является эффективным способом анализа текстовых данных, который помогает выявить основные темы и понять контекст использования ключевых слов. В нашем случае с помощью визуализации коллокаций мы увидели, что в основном объекты в нашей выборке связаны с политикой и документами (Рисунок 14).

Рисунок 14 - Визуализация в виде облака коллокаций, построенная по названиям объектов нашей выборки.

Изучение культурных данных требует тщательного анализа метаданных и понимание того, сколько объектов содержат пропущенные данные. Отсутствие данных может привести к ошибочным выводам и

гипотезам, а также существенно ограничить возможности дальнейшего исследования. Для решения данной проблемы был разработан метод визуализации данных, представляющий собой гистограмму с накоплением, отображающую процент пропущенных данных по каждому полю метаданных. При помощи этого графика можно установить, в каких полях метаданных недостаточно информации. Например, после построения гистограммы (Рисунок 15) было установлено, что количество объектов, с пропущенными описанием и автором, составляет значительную долю от общего числа, что подчеркивает необходимость тщательного анализа метаданных перед исследованием культурных данных. Таким образом, использование гистограммы с накоплением является эффективным инструментом для анализа культурных данных, позволяющим установить проблемы с отсутствием данных и, таким образом, более точно определить направления исследования в данной области.

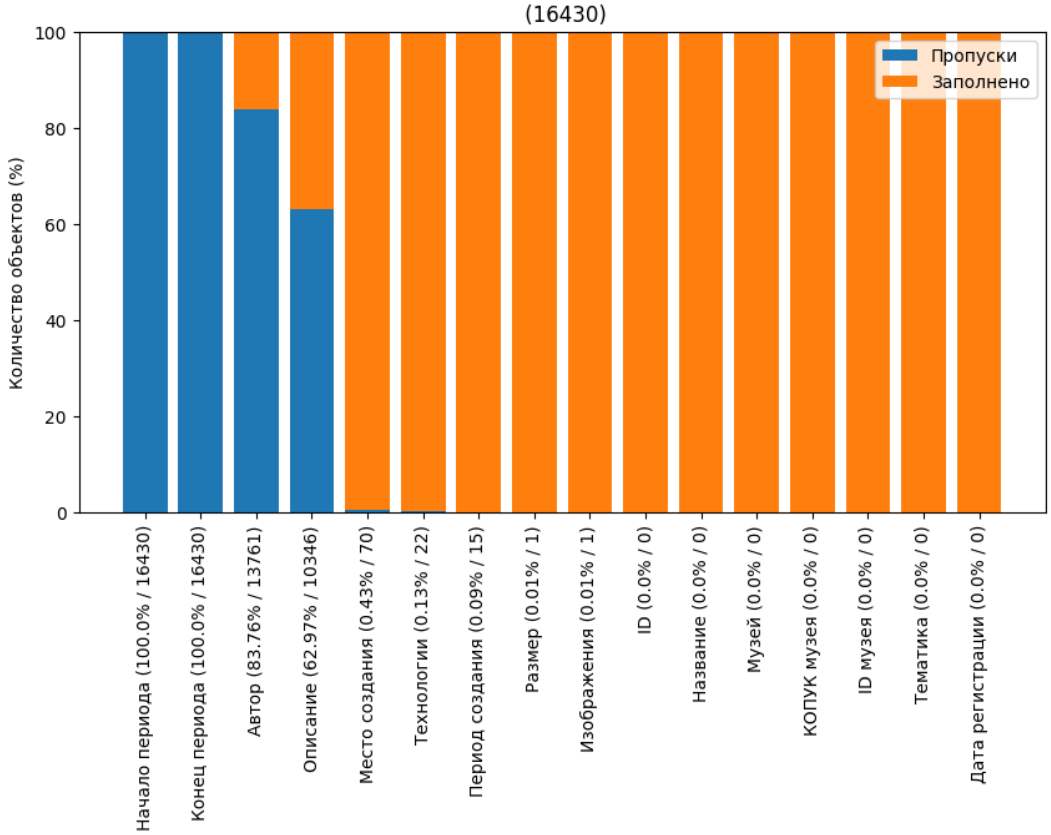

Рисунок 15 - Визуализация пропусков метаданных объектов нашей выборки в виде гистограммы с накоплением.

Последним способом визуализации программы SGAT является диаграмма, отображающая распределение объектов по времени. Важно знать, когда были созданы изучаемые объекты, чтобы формулировать гипотезы и делать выводы. Ранее была озвучена проблема с переводом нечетких дат в машиночитаемый формат, но также существует и проблема в выборе способа визуализации таких дат. Например, в своей докторской диссертации [23] Оливия Уэйн преобразовывала нечеткие даты в четкие, выбрав случайное число в пределах диапазона периода создания объекта. На наш взгляд, этот метод не идеален, поскольку он может привести к значительному неконтролируемому искажению графика. Мы решили эту проблему иначе. По оси X мы разместили шкалу времени, где минимальным отрезком будет считать один год. По оси Y откладывается количество объектов по каждому году. Четкие даты размещаются на графике стандартным способом, а именно увеличивают на единицу количество объектов в конкретном году. Нечеткую дату мы преобразовываем в массив, состоящий из всех годов промежутка этой даты (например, «1965-1970» будет преобразован в [1965, 1966, 1967, 1968, 1969, 1970]). Далее на линии времени к каждому году, входящему в промежуток заданной нечеткой даты, прибавляется не единица, а единица, деленная на количество годов в промежутке (например, при заданной нечеткой дате «1965-1970», график в каждом годе этого промежутка поднимется на 0,16). Таким образом, с помощью данного метода мы равномерно распределяем на линии времени нечеткую дату.

На построенном графике (Рисунок 16) можно заметить 3 построенные кривые. Кривая синего цвета показывает общее количество определенных четких и нечетких дат, в пределах заданного промежутка времени. Кривая зеленого цвета показывает только четкие даты, а кривая оранжевого цвета показывает только нечеткие даты. Так, данный график помогает увидеть распределение объектов нашей выборки по времени, а также оценить количество четких и нечетких дат.

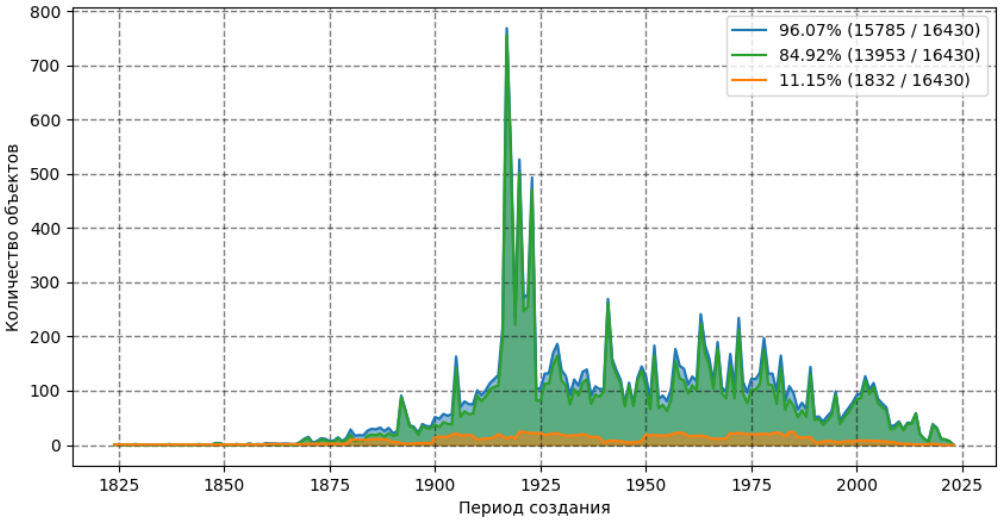

Рисунок 16 - Визуализация периодов создания объектов нашей выборки с 1823 по 2023 годы. На графике представлено распределение по времени 96% объектов.

При работе с большими объемами данных, визуализация является незаменимым инструментом, который помогает нам проанализировать информацию более эффективно. Путем отображения данных в более доступном и наглядном формате, мы можем быстрее обнаружить скрытые закономерности и взаимосвязи, которые могут быть незаметны в других форматах. Таким образом, визуализация имеет большое значение, поскольку позволяет представлять сложные данные в более простой форме и делать анализ более продуктивным и точным.

3 Примеры практического применения программы SGAT

3.1 Создание набора данных для обучения нейронной сети

Существует значительная проблема пропусков в метаданных, которая является критической для исследований. Данная проблема напрямую влияет на точность информации и может привести к неверным выводам. Попробовать решить эту проблему можно с помощью искусственного интеллекта. То есть необходимо разработать нейросеть, которая будет обучена определять часть полей метаданных на основе фотографии культурного объекта. Для обучения нейросети необходимо использовать качественный и большой датасет. В качестве источника данных мы выбрали Госкаталог, на котором хранится большое количество размеченных культурных данных.

В рамках этой работы мы решили собирать датасет для обучения нейросети из объектов археологической коллекции. Для этого было отобрано 26 музеев, где объекты имеют качественные фотографии и метаданные (Таблица 7). Это позволило сформировать датасет, который обладает достаточным количеством объектов и метаданных для обучения нейронной сети.

Таблица 7 – Отобранные музеи с качественными фотографии археологических объектов и правильно заполненными метаданными.

Ссылка на музей	**КОПУК музея**	**Количество объектов коллекции «предметы археологии»**
https://goskatalog.ru/portal/#/museums?id=15	159768	931
https://goskatalog.ru/portal/#/museums?id=61	159077	5534
https://goskatalog.ru/portal/#/museums?id=1392	111523	19088
https://goskatalog.ru/portal/#/museums?id=2872	150829	506
https://goskatalog.ru/portal/#/museums?id=2897	157527	1418
https://goskatalog.ru/portal/#/museums?id=4153	155324	4934
https://goskatalog.ru/portal/#/museums?id=2441	123056	16873
https://goskatalog.ru/portal/#/museums?id=1431	121050	95420
https://goskatalog.ru/portal/#/museums?id=631	114019	32194

https://goskatalog.ru/portal/#/museums?id=861	111237	29758
https://goskatalog.ru/portal/#/museums?id=1306	111186	86636
https://goskatalog.ru/portal/#/museums?id=3776	156435	690
https://goskatalog.ru/portal/#/museums?id=10329	152519	214
https://goskatalog.ru/portal/#/museums?id=4330	153210	4044
https://goskatalog.ru/portal/#/museums?id=650	114076	5359
https://goskatalog.ru/portal/#/museums?id=621	118151	27776
https://goskatalog.ru/portal/#/museums?id=530	127305	52247
https://goskatalog.ru/portal/#/museums?id=1299	110770	2814
https://goskatalog.ru/portal/#/museums?id=1518	125225	192
https://goskatalog.ru/portal/#/museums?id=1822	115238	206
https://goskatalog.ru/portal/#/museums?id=1970	133804	139
https://goskatalog.ru/portal/#/museums?id=2220	121652	54374
https://goskatalog.ru/portal/#/museums?id=1973	117200	107536
https://goskatalog.ru/portal/#/museums?id=2787	142432	10628
https://goskatalog.ru/portal/#/museums?id=95	110287	230473
https://goskatalog.ru/portal/#/museums?id=1936	133359	19851

Далее с помощью программы SGAT мы собрали датасет и построили круговую гистограмму, показывающую распределение объектов по признаку материала изготовления (Рисунок 17). Также стоит отметить, что отбирались объекты, имеющие только один материал изготовления в поле «технологии». В итоге, все объекты нашей выборки были разделены на 7 классов.

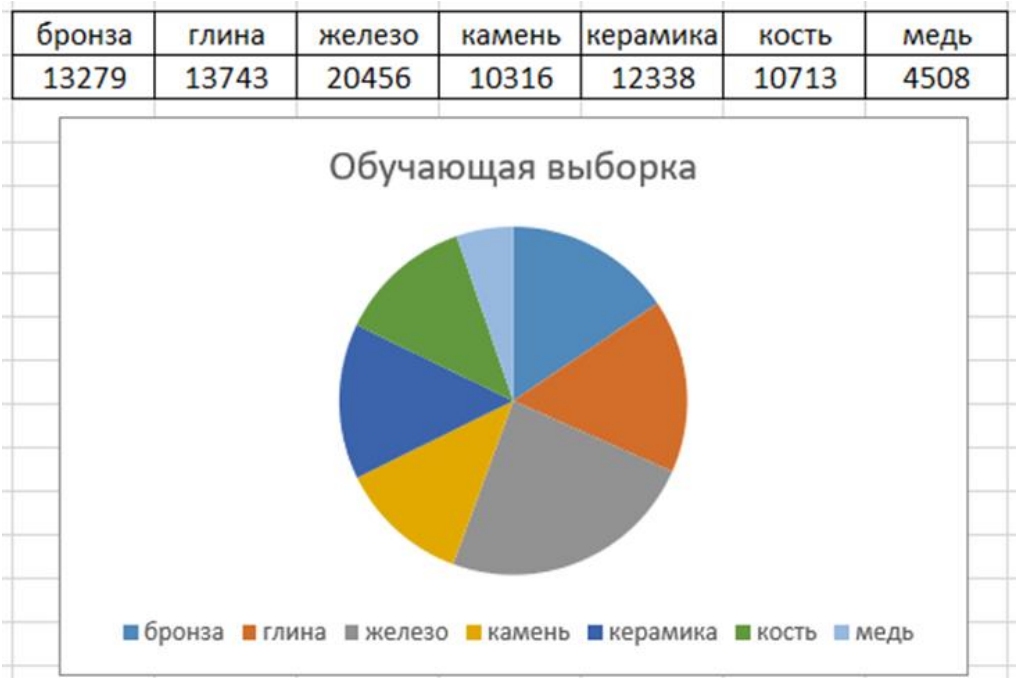

Рисунок 17 – Распределение материалов объектов в нашей выборке.

Далее с помощью алгоритма (Рисунок 18) были собраны изображения объектов Госкаталога. Скачивание каждого изображения занимало около 1-2 секунд. Данный алгоритм можно было бы запустить в многопоточном режиме, чтобы сократить время скачивания изображений, но мы столкнулись с тем, что сайт не мог обрабатывать одновременно несколько запросов с одного компьютера и отдавал «битое» изображение. Так, на данный момент, скачивание изображений с Госкаталога является времязатратной задачей.

```
import shutil
import requests
import os

name = input('Имя TXT: ')
folder_save = input('Место загрузки: ')

if not os.path.exists(folder_save):
    os.makedirs(folder_save)

column_o = []
try:
    with open(""+name+".txt", encoding="utf-8") as file: column_o = [row.strip() for row in file]
except FileNotFoundError:
    column_o = []

for i in range(len(column_o)):
    url = column_o[i]

    response = requests.get(url)
    rsp = response.content
    rsp = str(rsp)
    if rsp == "b''":
        print(url)
    else:
        response = requests.get(url, stream=True)
        with open(folder_save+'/'+url.split('/')[-1]+'.jpg', 'wb') as out_file:
            shutil.copyfileobj(response.raw, out_file)
        del response
```

Рисунок 18 – Алгоритм для скачивания изображений объектов Госкаталога.

Далее на основе нашего набора данных была обучена нейросеть с высокой точностью определения материала изготовления по фотографии археологического объекта (точность выше 90%). Таким образом, фотографии и метаданные объектов Госкаталога подходят для создания программных решений с участием искусственного интеллекта.

3.2. Анализ пропусков в метаданных объектов Госкаталога

Госкаталог представляет собой обширную коллекцию культурных объектов России, предоставляющую множество возможностей для

исследований. Однако, перед тем, как приступать к анализу данных, необходимо убедиться в их целостности.

С помощью алгоритмов программы SGAT нам удалось собрать информацию о более чем 34 миллионах объектов Госкаталога (94% от общего числа объектов). Далее с помощью модуля визуализации программы мы создали график, показывающий пропуски в метаданных объектов (Рисунок 19). Больше всего пропусков можно наблюдать в полях «автор» и «описание». Для некоторых исследований, например, таких как анализ художественных произведений, эти два поля могут быть максимально значимыми, поэтому перед исследованием важно проводить такой анализ данных.

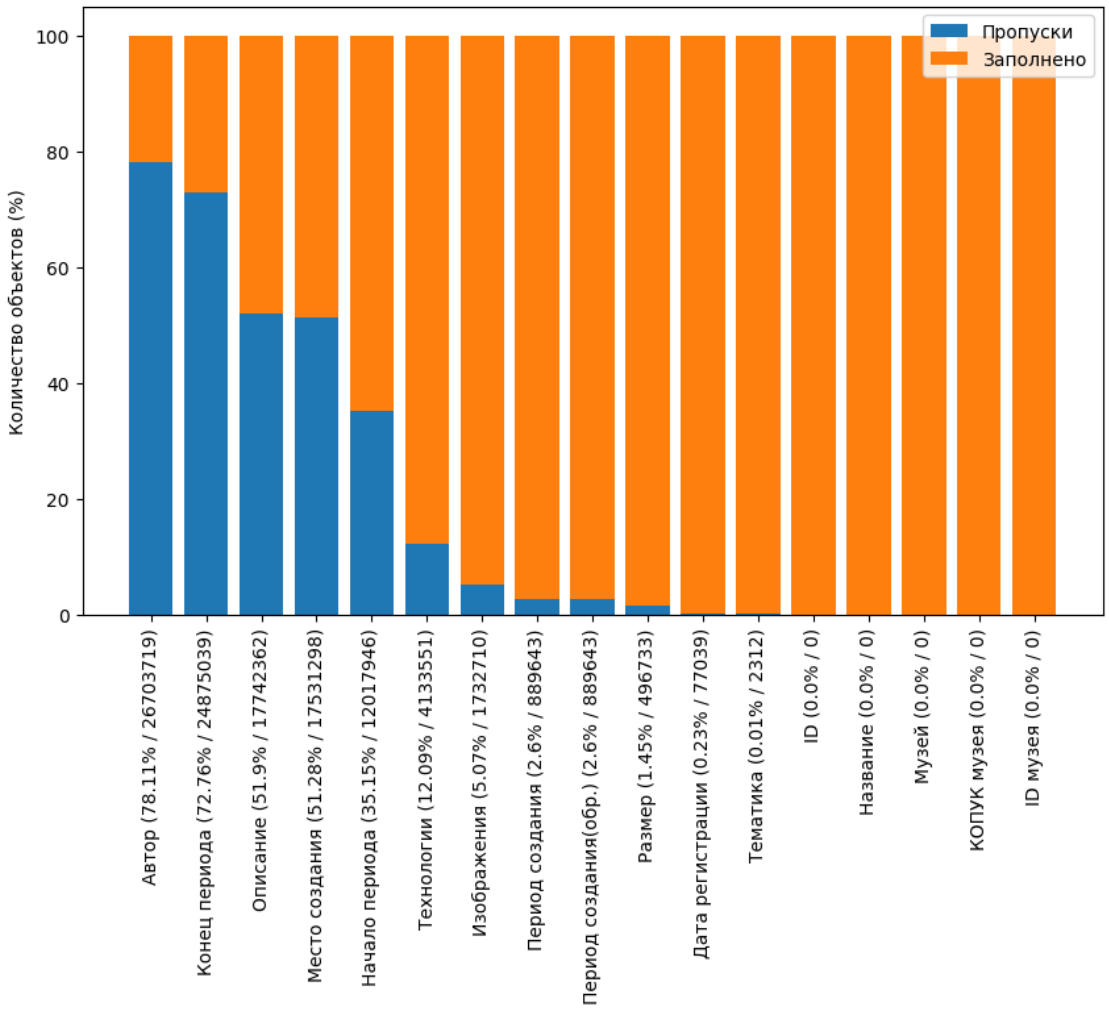

Рисунок 19 - Визуализация пропусков метаданных объектов Госкаталога в виде гистограммы с накоплением.

Далее было решено сконцентрироваться на меньшей выборке и изучить коллекцию живописи Госкаталога, которая насчитывает более 600 тысяч объектов. С помощью модулей предобработки и визуализации дат был сформирован график (Рисунок 20), который учитывает более 82% объектов живописи. На графике видно, что большое количество объектов было создано во второй половине 20 века. Эту информацию также важно учитывать перед проведением исследований, чтобы иметь представление о скосе в данных. Также можно заметить, что около 30% процентов объектов имеют нечеткую дату и без алгоритма визуализации нечетких дат программы SGAT было бы сложно учитывать эти данные на графике.

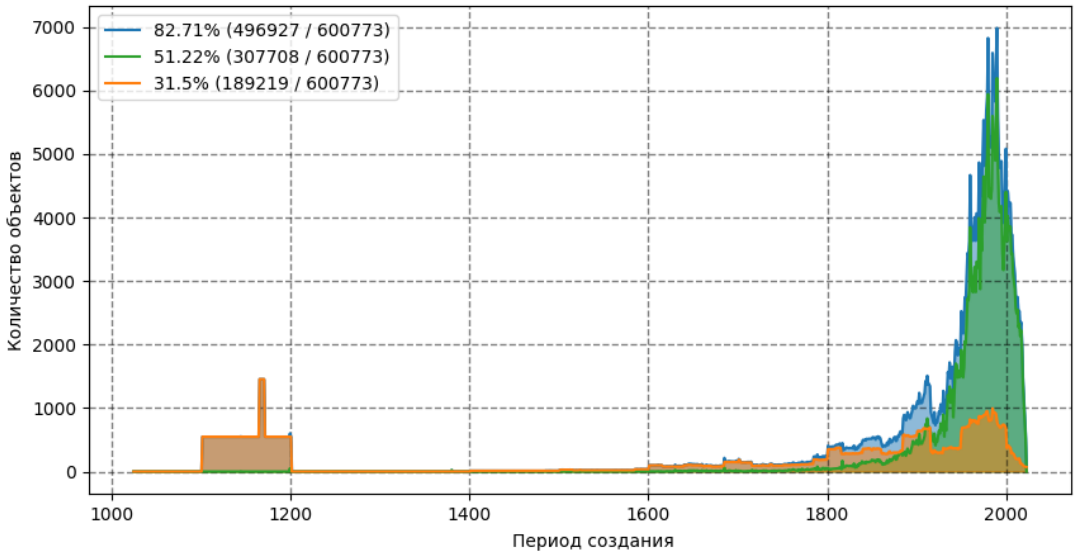

Рисунок 20 - Визуализация периодов создания объектов Госкаталога с 1023 по 2023 годы. На графике представлено распределение по времени 82,7% объектов.

Также с помощью параметров модуля визуализации программы SGAT можно подробнее рассмотреть распределение объектов живописи по времени с 19 по 20 век (Рисунок 21). Здесь еще больше виден скос в данных в сторону второй половины 20 века.

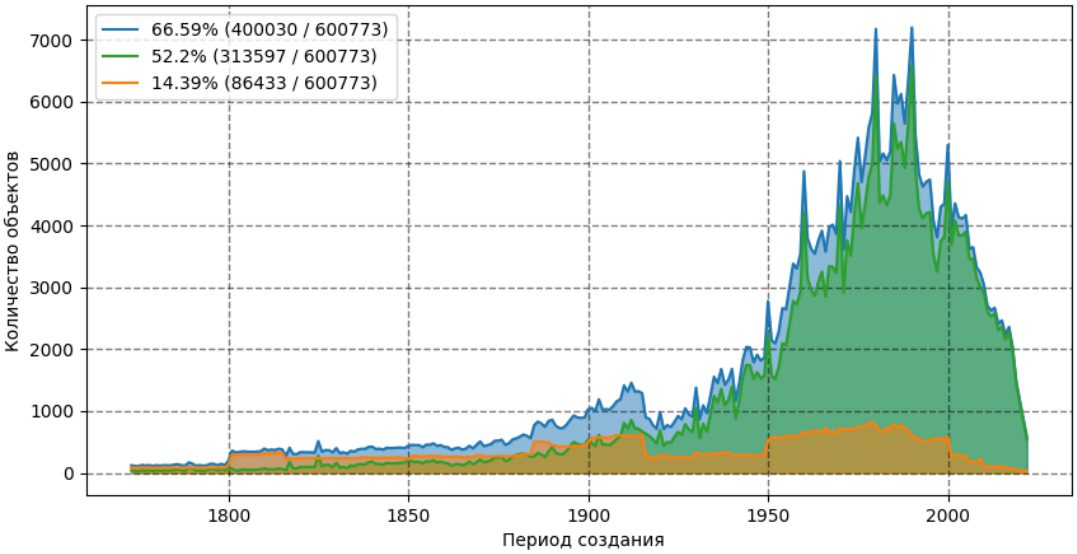

Рисунок 21 - Визуализация периодов создания объектов Госкаталога с 1173 по 2023 годы. На графике представлено распределение по времени 66,6% объектов.

Далее с помощью модулей программы можно будет визуализировать остальные коллекции Государственного каталога Музейного фонда РФ, чтобы выявлять пропуски в данных, закономерности и скосы. В настоящее время Госкаталог содержит около половины всех объектов Музейного фонда, поэтому представляет интерес выяснить, как изменятся графики, созданные по метаданным культурных объектов, когда коллекции будут полностью опубликованы.

ЗАКЛЮЧЕНИЕ

В данной работе была создана программа для сбора, обработки и визуализации метаданных культурных объектов SGAT. Программа основана на технологиях обработки естественного языка, веб-скрэйпинга, API, библиотеке языка программирования Python Matplotlib и других инструментах. С помощью программы исследователи в области цифровых гуманитарных наук смогут без труда использовать коллекцию Государственного каталога Музейного фонда РФ, проверять гипотезы, применяя встроенные модули обработки и визуализации данных, что, несомненно, важно, ведь это 37 миллионов данных о культурном наследии России. При этом работа демонстрирует, как можно использовать SGAT для исследований в области цифровых гуманитарных наук: создание обучающей выборки для нейронной сети определения материала объекта; изучение пропусков и распределения культурных данных коллекции Госкаталога. Не менее важно и то, что сейчас нельзя работать со всеми данными коллекции Госкаталога, не создав программный код. Программа, в свою очередь, дает не только возможность изучать культурные данные Госкаталога, но и сразу сделать визуализацию своих результатов. Так, например, в программу внедрена уникальная технология визуализации дат, которые представлены промежутком, что решает проблему скоса в результатах анализа дат, где есть и промежутки, и даты обычного формата.

Далее планируется улучшать алгоритмы обработки и визуализации, добавлять новые компоненты. Также программа SGAT является хорошим инструментом-фундаментом, с помощью которого мы планируем решить проблему стандартизации данных, пропусков данных, а также получать новые знания в сфере цифровых гуманитарных наук.

Программа опубликована и находится в открытом доступе на github, поэтому каждый исследователь, желающий работать с культурными данными Госкаталога, может беспрепятственно ей воспользоваться.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Europeana [Электронный ресурс] – Режим доступа: https://www.europeana.eu/en – Дата доступа: 2023.
2. The Metropolitan Museum of Art [Электронный ресурс] – Режим доступа: https://www.metmuseum.org/ – Дата доступа: 2023.
3. Freire N., Isaac A. Wikidata's linked data for cultural heritage digital resources: An evaluation based on the Europeana Data Model //International Conference on Dublin Core and Metadata Applications. – 2019. – С. 59-68.
4. Kouretsis A. et al. Mapping Art to a Knowledge Graph: Using Data for Exploring the Relations among Visual Objects in Renaissance Art //Future Internet. – 2022. – Т. 14. – №. 7. – С. 206.
5. Raemy J. A. Enabling better aggregation and discovery of cultural heritage content for Europeana and its partner institutions : дис. – Haute école de gestion de Genève, 2020.
6. Kaldeli E. et al. Europeana Translate: Providing multilingual access to digital cultural heritage //Proceedings of the 23rd Annual Conference of the European Association for Machine Translation. – 2022. – С. 299-300.
7. Villaespesa E., Crider S. Computer Vision Tagging the Metropolitan Museum of Art's Collection: A Comparison of Three Systems //Journal on Computing and Cultural Heritage (JOCCH). – 2021. – Т. 14. – №. 3. – С. 1-17.
8. Zhitomirsky-Geffet M., Kizhner I., Minster S. What do they make us see: a comparative study of cultural bias in online databases of two large museums //Journal of Documentation. – 2023. – Т. 79. – №. 2. – С. 320-340.
9. Villaespesa E., Crider S. A critical comparison analysis between human and machine-generated tags for the Metropolitan Museum of Art's collection //Journal of Documentation. – 2021.
10. Государственный каталог музейного фонда [Электронный ресурс] – Режим доступа: https://goskatalog.ru/portal/ – Дата доступа: 2023.
11. Академия Google [Электронный ресурс] – Режим доступа: https://scholar.google.ru/ – Дата доступа: 2023.

12. Kizhner I. et al. The Culture of the Very Rich and Very Poor: Do Digital Museum Collections Tell us Anything about Jewish Culture? //Studies in Digital History and Hermeneutics. – 2022. – Т. 43.

13. Портал открытых данных Минкультуры России [Электронный ресурс] – Режим доступа: https://opendata.mkrf.ru/ – Дата доступа: 2023.

14. Глазунов Е. В. и др. УНИФИКАЦИЯ ДАННЫХ МУЗЕЙНОГО ГОСКАТАЛОГА РФ //Сибирский антропологический журнал. – 2020. – Т. 4. – №. 3. – С. 154-168.

15. Кожин К.Д. Программа для скрэйпинга и анализа открытых данных культурного наследия (SGAT). Свидетельство №2022680022 о гос. регистрации в Реестре программ для ЭВМ от 03.11.2022.

16. Репозиторий программы SGAT [Электронный ресурс] – Режим доступа: https://github.com/konstantinkozhin/SGAT – Дата доступа: 2023.

17. Requests [Электронный ресурс] – Режим доступа: https://requests.readthedocs.io/en/latest/ – Дата доступа: 2023.

18. Selenium [Электронный ресурс] – Режим доступа: http://selenium.dev – Дата доступа: 2023.

19. Pullenti [Электронный ресурс] – Режим доступа: https://pullenti.ru/ – Дата доступа: 2023.

20. Natural Language Toolkit [Электронный ресурс] – Режим доступа: https://www.nltk.org/ – Дата доступа: 2023.

21. Pymorphy2 [Электронный ресурс] – Режим доступа: https://pymorphy2.readthedocs.io/en/stable/ – Дата доступа: 2023.

22. Matplotlib [Электронный ресурс] – Режим доступа: https://matplotlib.org/ – Дата доступа: 2023.

23. Vane, O. Timeline design for visualising cultural heritage data. PhD dissertation. Royal College of Art, UK, 2019. https://www.oliviavane.co.uk/phd

Министерство науки и высшего образования РФ
Федеральное государственное автономное
образовательное учреждение высшего образования
«**СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ**»

Гуманитарный институт
Кафедра информационных технологий
в креативных и культурных индустриях

УТВЕРЖДАЮ
И. о. заведующего кафедрой
_____ М. А. Лаптева
«_____» _____ 2023 г.

БАКАЛАВРСКАЯ РАБОТА

Система автоматизированного получения и анализа открытых данных культурного наследия.

Направление подготовки: 09.03.03 Прикладная информатика

Наименование программы: 09.03.03.30 Прикладная информатика

Руководитель проф., д-р тех. наук О. А. Антамошкин

Выпускник К. Д. Кожин

Нормоконтролер И. Р. Нигматуллин

Красноярск 2023