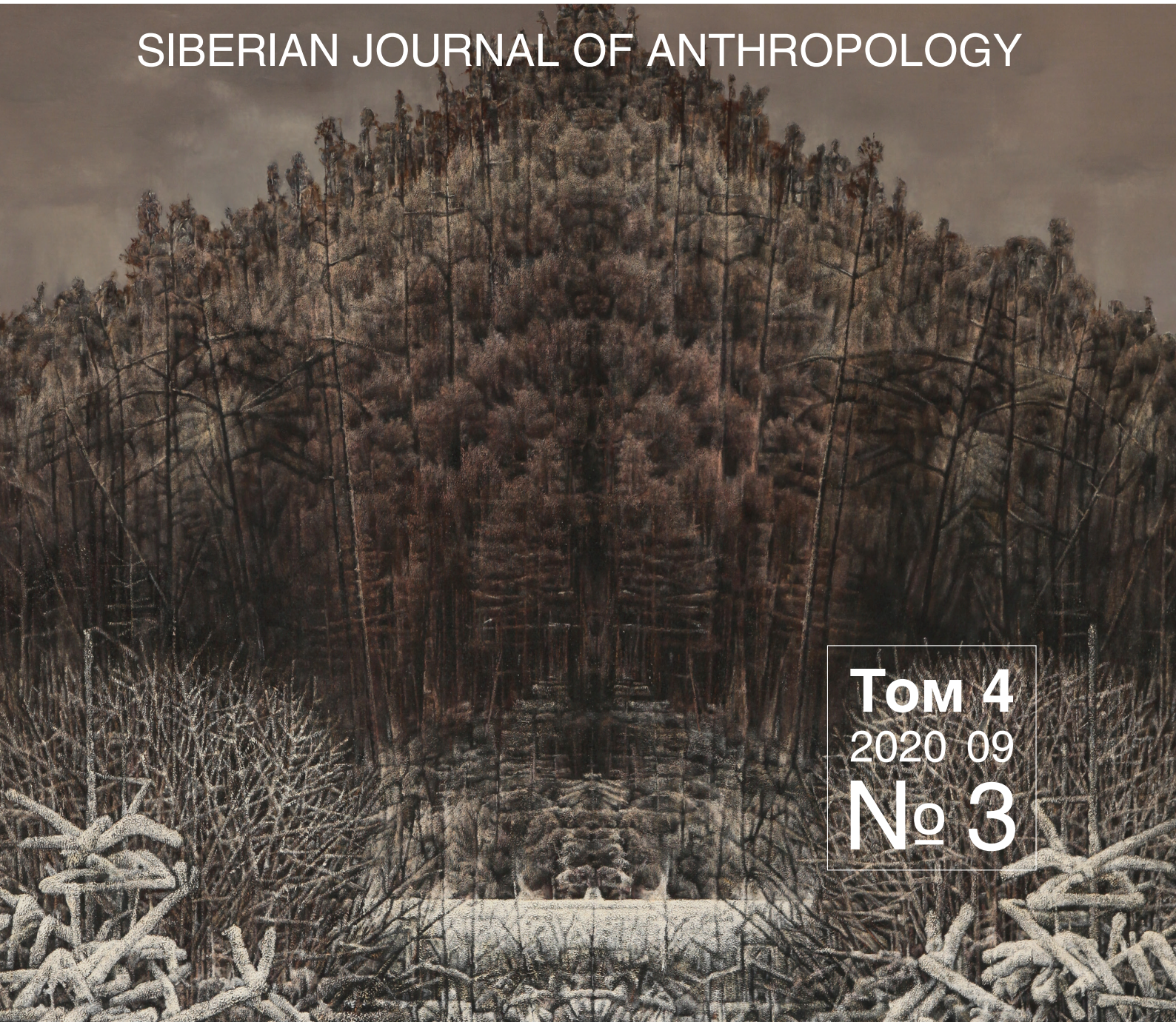
2020 09 (4/3)

ISSN 2542-1816

СИБИРСКИЙ АНТРОПОЛОГИЧЕСКИЙ ЖУРНАЛ

SIBERIAN JOURNAL OF ANTHROPOLOGY



Том 4
2020 09
№ 3

СОДЕРЖАНИЕ

CONTENTS

Раздел 2.
АНТРОПОЛОГИЯ ИСКУССТВА

Section 2.
ART ANTHROPOLOGY

Раздел 3.
ЦИФРОВОЙ МИР КУЛЬТУРЫ

Раздел 4.
ИСТОРИЧЕСКАЯ АНТРОПОЛОГИЯ

Section 3.
DIGITAL WORLD OF CULTURE

Section 4.
HISTORICAL ANTHROPOLOGY

Раздел 3. Цифровой мир культуры

УДК 069.013

УНИФИКАЦИЯ ДАННЫХ МУЗЕЙНОГО ГОСКАТАЛОГА РФ

Глазунов Евгений Владимирович[1]
Орехов Борис Валерьевич[2]
Национальный исследовательский университет
«Высшая школа экономики»

Аннотация

В эпоху больших данных растет интерес в том числе и к большим данным в гуманитарных сферах (например, в рамках цифровой гуманитаристики). В России существует Государственный каталог Музейного фонда Российской Федерации, где собирается информация об экспонатах из коллекций российских музеев. В настоящее время описано более 16 млн экспонатов. Многие поля в базе заполнены на естественном языке, например, «XIX век» в качестве даты создания предмета, что делает практически невозможным анализ данных. Инструменты автоматической обработки языка (например, извлечение именованных сущностей) позволяют унифицировать данные и привести их в удобный для анализа вид. В работе показано, как можно обрабатывать информацию о дате создания, месте создания, авторах и техниках. В качестве иллюстрации того, как можно использовать нормализованные данные, приводится некоторая аналитика по странам и периодам в разных категориях экспонатов, что позволяет увидеть известные закономерности.

Ключевые слова: *цифровые гуманитарные науки, Госкаталог, базы данных.*

PROCESSING DATA OF THE RUSSIAN MUSEUM STATE CATALOGUE

Glazunov Evgenii Vladimirovich[1]
Orekhov Boris Valerjevich[2]
National Research University
"Higher School of Economics"

Abstract

In the era of big data the interest in big data in humanities is growing (for example, in digital humanities). In Russia there is the Russian Museum State Catalogue that contains information about objects in Russian museums. The work is still in progress but it already contains information about more than 16 million objects. A lot of data fields are written in natural language and it makes data analysis almost impossible. Instruments of natural language processing (for example, named-entity recognition) help to process data and make it possible to analyse it. In this

[1] © Glazunov E. V., 2020. Corresponding author Email: e.glznv@yandex.ru

[2] © Orekhov B. V., 2020. Corresponding author Email: nevmenandr@gmail.com.

work we describe the processing of the date of creation, the place of creation, authors and used techniques. As an example of a research on processed data we describe different categories of objects from the perspective of place of origin and time of creation. The results meet expectations (well-known facts about history of art).

Keywords: *Digital Humanities, databases.*

Научная специальность: *24.00.03 – Музееведение, консервация и реставрация историко-культурных объектов (искусствоведение).*

Doi: *10.31804/2542-1816-2020-4-3-154-168*

1 Введение

Интерес к цифровизации и большим данным, в том числе и в области изучения культуры, растет. Как продолжение этого интереса возникла область, которая изучает большие датасеты оцифрованных культурных объектов с помощью программных инструментов и восстанавливает скрытые паттерны в таких данных (Said, 2000; Deuze, 2006). Традиционно исследователи-гуманитарии работают с относительно небольшими наборами данных и проводят преимущественно качественный анализ (Van Peer и др., 2012). Специалисты в области компьютерных наук, с другой стороны, способны обрабатывать большие данные, применять статистический анализ и сложные техники моделирования, преследуя, как правило, практические цели, не свойственные гуманитарным наукам, например, предсказание поведения пользователя. Использование этих инструментов для исследования культуры приносит новые перспективы. Основная среди них – это возможность представить культурные объекты в многомерном пространстве признаков (географических, хронологических, визуальных), для чего можно использовать не только традиционное приписывание объектам меток с помощью ручного труда, но и автоматическое извлечение признаков. Эти многомерные представления помогут взглянуть на объекты качественно новым образом и одновременно составить представления о целых культурных эпохах, основываясь при этом на измеримых показателях.

Министерство культуры РФ развивает инициативу машиночитаемой базы музейных экспонатов Госкаталог (Государственный каталог Музейного фонда Российской Федерации) , в которую потенциально планируется внести все хранимые объекты. Аналогичная база существует во Франции: Joconde Database of French Museum Collections , а также есть примеры цифровых коллекций отдельных музеев (Metropolitan Museum of Art в США, Rijksmuseum в Нидерландах). Российский Госкаталог пока не завершен, однако там уже находится более 16 миллионов записей. Учитывая объем данных, это тот случай, когда для обработки данных сложные компьютерные инструменты и методы культурной аналитики необходимы. Пока мы видим не так много примеров использования этого набора данных в научной сфере, однако интерес к нему растет вместе с повышением уровня необходимых для этого навыков исследователей. Выгода от работы с этими данными очевидна не только для академической среды, но и для самих музеев. Востребованность данных Госкаталога за пределами сообщества музейных работник способна дать им дополнительную мотивацию для заполнения базы.

2. Исследовательские вопросы

В настоящей работе мы ставим себе цель унификации части данных Госкаталога. На текущем этапе это даты, место создания, авторы, техники и материалы. Унификация необходима для того, чтобы данные из различных источников (музеев) были сопоставимы. Без достижения этой цели невозможно обратиться к собственно исследовательским вопросам, касающимся структуры коллекции в россий-

ских музеях. Эта структура может быть описана через ответы на более частные вопросы:

• Из каких стран пришли музейные экспонаты? Есть ли статистические различия в их описании в зависимости от времени создания?

• Какие наиболее популярные техники и материалы используются для создания музейных экспонатов, хранимых в России? Какие из этих техник чаще фигурируют в описаниях вместе?

• Какие периоды, страны и категории объектов (такие как живопись, скульптура, техника, оружие) наиболее представлены в Госкаталоге?

3. Данные

Данные, с которыми мы работали, представляют собой копию базы, доступной онлайн на сайте открытых данных Министерства культуры России. База регулярно обновляется, в этой работе используется версия, вывешенная в марте 2020 года. База состоит из набора JSON-файлов, содержащих информацию о музейных объектах, которую можно подразделить на:

• техническую информацию о времени обновления записи,

• уникальный идентификатор, категорию объекта (скульптура, живопись, документ и т.д.),

• место создания,

• автор,

• дата создания,

• идентификатор музея,

• техники и материалы, размер и некоторую другую информацию

Главная проблема для машинной обработки заключается в том, что ключевые поля заполнены на естественном языке без следования требованиям к формату записи или с ошибками в существующем формате. Это не составляет сложности для человека, однако затрудняет любую автоматизированную работу. При таком разнобое в данных все еще остается возможным поиск (он реализован на сайте проекта), однако аналитика становится практически недоступной: у пользователя нет возможности задать запрос к базе, который помог бы получить в результате список всех работ одного автора или периода. Дело в том, что имя автора может быть отображено в базе многими способами («Ф. П. Толстой», «Толстой Ф.»), а даты представлены на порядок более разнообразно («40-50-е гг 20 века», «1940-е» и т.д.). Эта проблема не специфична для записей об авторах или датах, и касается многих полей в базе.

На сайте Министерства культуры также существуют и другие базы, относящиеся к интересующей нас сфере, они содержат статистику музейных учреждений и их данные могут быть привязаны к Госкаталогу с помощью идентификаторов.

4. Унификация

Унификация подразумевает трансформацию представленных на естественном языке данных в структурированный машиночитаемый формат. В этой работе сосредоточились на 4 полях базы данных: дата создания, техники и материалы, место создания и авторы.

4.1 Даты

Данные Госкаталога содержат информацию о времени создания экспоната на естественном языке. Некоторые записи имеют в дополнение к этому и формализованный вид. Не ясно, кто вносил формализованную информацию. Нельзя исключать, что этим занимались не те же сотрудники, которые заполняли сведения на естественном языке, так как можно наблюдать различия или отдельные несовпадения формализованного и текстового варианта («1907 г., 7октября» – > «1896-01-01» по «1906-01-01»). Проблема текстового представления заключается в большой вариативности записей:

• вариативность знаков препинания: точки, запятые, тире, дефисы («1971 г. Для бланка – 1971 г.», «[Начало XX в.]»)

• замена букв на графически похожие в римских цифрах (смешение латинского, русского и греческого алфавитов) и в окончаниях – x, -е («1990-х гг», «40-е»)

• склеивание цифр и следующих слов, произвольные пробелы в составных частях («1990 х гг», что мешает снимать омонимию вида век римскими цифрами или указание периода годов)

• конструкции вида «40-е гг 20 века» наряду с «1940-е гг»

• отсутствие указания века для XX века, что не позволяет отличить такие даты от I века н. э. (в результате в базе фигурируют фотографии I в. н. э.)

• словесное обозначение временного периода («средние века», «античность», «палеолит»).

• неформальные определения временного диапазона вроде «конец, начало, середина», которые не имеют строгой интерпретации.

• отсутствие стандарта для пропущенных данных («б.д.», «не установлено» и пр.), что мешает, в частности, оценке заполненности базы.

• опечатки («20-18 ввв. до н.э.», «03 мачя 1943 год»)

Как уже было сказано, формализованные даты не всегда точны. Нельзя игнорировать и проблемы, связанные со способом их хранения. В Госкаталоге даты хранятся в виде 0000-00-00, что ограничивает диапазон доступных значений от 0001-01-01 – 9999-12-31. Он покрывает подавляющее большинство экспонатов, но не учитывает потребности описания объектов, созданных до нашей эры. Можно наблюдать, как музейные работники при заполнении каталога стараются выйти из положения. Очевидно, что если в записи диапазона значение первой даты больше, чем второй даты, то речь идет о датировке до нашей эры. Однако такие недокументированные способы фиксации создают омонимию дат. Например, диапазон «0100-01-01 – 0200-01-01» может значить как период с I века до н. э. по II век н. э., так и II век н. э. Из этого следует необходимость хранить сведения об эре. С нашей точки зрения, полезным было хранить даты в 3 разных полях: год, месяц, день. Такой способ позволяет учесть, что последние два числа имеют фиксированный небольшой диапазон и оптимизировать под них память. Кроме того, даты до н. э. можно было бы обозначать как отрицательное значение первого поля (-2000).

Компьютерная лингвистика предоставляет инструменты для того, чтобы автоматизировать работу с датами, главным образом, их извлечение из текста на естественном языке. Они основаны на разных принципах (правила, машинное обучение, глубокое обучение), однако без модификаций под конкретную задачу имеют свои недостатки. На данных каталога многие даты не распознавались (например, века римскими цифрами).

В этой работе решено использовать правиловый (rule based) подход, который кажется оправданным на первых этапах работы по нескольким причинам. Во-первых, мы имеем дело с довольно узкой задачей: поиск даты производится не в произвольном тексте, а в специализированном поле базы, которое редко содержит лишнюю информацию. Подходы, основанные на машинном обучении, эффективнее показывают себя в обратных случаях. Во-вторых, доработка парсера на правилах проще, чем обучение моделей, и не требует большой выборки. Подход, основанный на машинном обучении, предполагал бы объемную ручную подготовку данных.

Для того, чтобы решить задачу распознавания временных диапазонов и дат, необходимо решить две задачи:

1. Промежуточная: выделение дат (дат может быть несколько, может содержаться информация, не относящаяся непосредственно к дате) .

2. Финальная: интерпретация дат и приведение их к формальному представлению.

Так как правила в случае подобной вариативности написаний могут быть очень сложными (особенно для редких версий), было принято решение исходить из соображений упрощения и приоритезации. Во-первых, решать промежуточную задачу в целом, а финальную в нескольких вариантах приближения (с учетом качества). Упрощение заключалось в сильной предобработке данных таким образом, чтобы добиться сокращения числа уникальных вариантов. Например, были нормализованы пробелы и знаки препинания. Приоритезация состояла в ранжировании правил от наиболее типичных к более редким,

Таблица 1.
Примеры шаблонов для поиска дат

XX в., первая половина	кон.19 – нач.20 вв.	июнь 1961	1980 г.	Палеолит. 25 тыс. лет т.н.
X в WNUM половина	конец 00 начало 00 в	имя_месяца 0000	0000 г	00 тыс т н

чтобы написанными N правилами покрыть наибольшую возможную часть базы.

Нормализация

Нормализация частично решает описанные проблемы и включает несколько этапов:

1. Нормализация неверных римский цифр (замена омографических русских или греческих букв на соответствующие латинские, где это возможно)
2. Разделение склеенных токенов («Хвек», «3я»)
3. Замена частых сокращений на полные варианты («сер» – «середина»)
4. Исключение слов, кроме числительных или аналогичных прилагательных и отобранного списка слов, которые являются частью даты («Учрежден Указом Президиума Верховного Совета СССР от 20.05.1942 г.»)

Для выделения паттернов будущих правил из нормализованных вариантов дат были получены маски (шаблоны), в которых числа заменены на условные обозначения, а текстовые части были сохранены (см. таблицу 1).

Такое выделение шаблонов позволяет сфокусироваться на самых популярных способах представления даты. Подсчеты показали, что 200 самых частотных шаблонов покрывают около 95 % всех записей базы. Общее количество таких шаблонов превышает 10 000, что, однако существенно меньше, чем уникальных записей в поле дата (около 807 000).

Составление целевого и контрольного набора данных

Из самых распространенных шаблонов написания дат нами получен набор для составления правил и последующей оценки качества нашего инструмента. Каждый объект набора имеет описание того, как первая часть (парсер для выделения) и вторая часть (интерпретатор) должны срабатывать на таких шаблонах. В качестве тестового примера выбран один из реальных вариантов из базы, который соответствует конкретному шаблону. Также там указан вес, который рассчитывается как общее число вхождений текстов, подходящих под такой шаблон:


```
{
    "id": 0, "weight": 4445645, "type": "0000 г",
    "text": "1980 г.", "text_process": "1980 г",
    "left": {"year": 1980, "month": 1, "day": 1, "bc": false},
    "right": { "year": 1980, "month": 12, "day": 31, "bc": false},
    "data": [
      {
        "roman": null, "integer_ct": null, "year": 1980, "month": null, "day": null,
        "current_era": true, "part": null, "season": null,
        "is_million": null, "is_thousand": null, "is_century": null, "is_year": true,
        "year_ago": null, "period": false
      }
    ]
  }
```

Ключ data относится к блоку выделения даты (или ее частей если дата составная), a left и right содержат итоговый результат. Качество на таком наборе рассчитывается как доля суммы правильных случае от всех случаев в базе и от суммы тех, что входят в тестовый набор. Это оставляет некоторую неопределенность и не позволяет точно оценить правильность срабатываний (вероятны ошибки при составлении шаблона и строки, двухразрядные числа могут быть и обозначением года, и века: 20 в.? г.?). Но приблизительную оценку таким способом мы получить сможем. Стоит отметить, что в какой-то степени этот набор является и обучающим, и тестовым, но в данном случае это не представляет проблемы в отличие от систем, основанных на машинном обучении.

Создание правил

Первый этап заключается в написании правил для извлечения и сортировки частей, входящих в многосоставное представление дат. Для этого используется инструмент создания правил yargy (пакет для Python), который применяется при извлечении именованных сущностей. Yargy состоит из частей, которые реализуют поиск по строке для нахождения желаемых последовательностей элементов шаблона. На основе yargy построена библиотека natasha, способная распознавать сущности и содержащая готовый набор правил для дат, имен, географических объектов. Однако этих правил недостаточно при работе со специфическими данными, к которым относятся и текстовые описания Госкаталога. Нам потребовалось создание собственных правил.

За основу нами были взяты готовые правила, подходящие под наши цели, а также составлены дополнительные, которые описывают типичные шаблоны, сформированные на предыдущем этапе. Парсер, к которому подключены эти правила, выделяет следующие части дат: римское число, числовое представление века или иное число, год, месяц, день, до нашей эры, часть периода («конец»), сезон («лето»), миллион, тысячелетие, век, количество лет назад, целый период (например, «1990-е» vs «1990»). Эти части представления дат в своей комбинации позволяют на следующем шаге интерпретировать дату как конкретные день, месяц и год и переводить их в формализованное представление.

Интерпретация дат

Интерпретация дат – это представление в формализованном виде числа, полученного на предыдущем шаге. В общих чертах можно описать алгоритм следующим образом (под «левой частью» понимается начало временного диапазона, под правой его конец):

1. Если дата односоставная, то левая часть копируется в правую (начало = конец)
2. Даты обмениваются общей информацией, если она только в одной части, и происходит общая обработка:
   a. интерпретируются римские цифры
   b. если правая часть до нашей эры, то эта пометка копируется в левую
   c. если правая часть тысячи, то и левая тоже
   d. если есть пометка «до», то левая часть убирается, если «после», то правая
   e. если «лет назад», то тысячелетия уменьшаются на 2 f. если оба года есть и первый четырехзначный, а правый двузначный, то второй превращается в тот же век (1870-80 – > 1870-1880)
   g. Если год меньше 100 и при это есть указание века, то копируется век из правого в левый (40-50 гг 20 века)
3. Преобразуется левая часть
   a. Обрабатываются случай раздельного года и века
   b. Прописываются стартовые даты для века или тысячелетия (см. правила далее) с учетом того, к какой эре относится
   c. Если одиночные числа 19 или 20, то они интерпретируются как век
   d. Если нет месяца или года, то ставятся по умолчанию (первое число для месяца, 1 января для года)

Таблица 2.
Интерпретация нечисловых описаний временных диапазонов

часть \ масштаб	век	тысячелетие	«десятые»
-	1901-2000	1001-2000	1910-1919
начало	1901-1920	1001-1200	1910-1912
середина	1930-1970	1300-1700	1913-1917
конец	1980-2000	1800-2000	1918-1919
первая половина	1901-1950	1001-1500	1910-1914
вторая половина	1951-2000	1501-2000	1915-1919
первая треть	1901-1933	1001-1333	1910-1913
вторая треть	1934-1966	1333-1667	1914-1916
последняя треть	1967-2000	1668-2000	1917-1919
четверть	1901-1925 1926-1950 1951-1975 1976-2000	1001-1250 1251-1500 1501-1750 1751-2000	1910-1912 1912-1914 1915-1917 1917-1919
рубеж	1890-1910	0900-1100	1909-1911
около	1801-2100	0001-3000	1900-1930

Для года “около” + – 5 лет

e. Если до нашей эры, то год умножается на – 1 4. Преобразуется правая часть
a. (Аналогично левой + минимальные корректировки)

Здесь интерпретируются только конкретные даты без учета случаев типа «начало», «рубеж», однако для формализации была разработана система соответствий, представленная в таблице 2.

Результат

К сожалению, интерпретация веков и тысячелетий пока не реализована, однако использованные для формализации шаблоны покрывают большую часть базы. Для того, чтобы сократить число ошибок, успешно интерпретированными считались только те даты, которые проходили проверку по контрольному набору. В рамках этой работы достаточным результатом считалось определение датировки с точностью до века.

Всего в экспонатов: 16039530

Имеющие информацию о дате: 14927438 (93 % от всех)

Среди них:

Таблица 3.
Заполненность формализованных дат до и после обработки. Проценты представлены относительно тех экспонатов, у которых заполнено текстовое представление даты.

	Год начала	Год окончания	Век начала	Век окончания
Исходные данные	10338886 69 %	4159974 27 %	нет такой категории	нет такой категории
Обработанные данные	11953482 80 %	11954504 80 %	13547661 91 %	13549769 91 %

Какие записи остаются неинтерпретированными? Под шаблоны не подпадают случаи типа «конец 19 века», аналогичные вхождения с тысячелетиями, часть указаний на рубеж эпох. К тому же существует много записей, указывающих на отсутствие даты («б.д.», «не указано»). Таким образом, среди записей, в которых дата указана, процент заполнения выше. В будущем необходимо добавить распознавание словесных обозначений периодов («средние века», «палеолит» и прочие).

4.2 География

Место создания музейного объекта, как и дата, указано на естественном языке и не имеет последовательного формата. В то же время в большинстве случаев можно угадать более-менее общий шаблон, согласно которому место указывается последовательно иерархически: от крупных к более локальным уровням, например, страна-регион-город, причем эти уровни разделены запятой.

Существует несколько возможных вариантов работы с геоданными. Во-первых, обращение к API Google или Yandex картам. Эти API позволяют адресовать к ним запрос с наименованием местности и в автоматическом режиме получать ответ с указанием широты и долготы. Этот способ позволил бы достичь хорошего качества распознавания местности, так как инженеры крупных картографических сервисов уже позаботились о многих нюансах распознавания мест. Однако большое число запросов требует использования платного тарифа, а бесплатные лимиты не подходят для масштабного проекта. Другая проблема состоит в специфике исторических данных. Современные карты ориентированы на актуальные названия стран, населенных пунктов и пр., а в записях Госкаталога зачастую используются исторические варианты (например, СССР). Вторым способом, бесплатным и относительно удобным, является онтология Wikidata и вообще экосистема Wikipedia. Эта инфраструктура включает интеллектуальный поиск, умеющий исправлять разные формы написания, и формализованное представление географических координат. Лимит на запросы к этой системе отстутствует.

Был применен следующий алгоритм работы с названиями мест создания экспонатов:

1. Нормализация текста и разделение по запятым на разные уровни иерархии.
2. Поиск каждой части в Wikipedia
3. Поиск каждого результата в Wikidata
4. Фильтрация классов мест
5. Выбор результата из вариантов
6. Определение географического положения
7. Проверка единообразия

Рассмотрим эти этапы подробнее.

Нормализация текста

Нормализация текста требовалась минимальная.

1) исправление недостающих пробелов между сокращениями и названиями («гМосква», «г.Москва»)

2) удаление квадратных скобок («[г. Саранск]»)

3) удаление информации в круглых скобках («Бельгия, г Льеж, фирма "J.B.ROGE FILS" (Роже и сыновья)»)

4) исправление точек или точек с запятой на запятые в случаях, когда они очевидно взаимозаменяемы («Тамбов; Электро-Типография Губернского Правления»)

5) добавление запятых после известных стран, чтобы отделить их от городов («Франция Париж»)

6) удаление множественных пробелов («Республика Дагестан»)

7) нормализация сокращений области и губернии, сокращенного названия Москвы («М.: Правда»)

Поиск по Wikipedia

Нормализованный текст кладется в основу списка уникальных названий, они сортируются по частоте для приоритезации дальнейшего поиска. Каждое из названий запрашивается в Wikipedia через библиотеку, реализующую

на Python работу с API Wikipedia. Для каждого места находится 10 кандидатов (топ выдачи поиска). Затем эти кандидаты сортируются по убыванию частотности (сумма оригинальных вхождений) и сохраняются в промежуточную таблицу.

Поиск по Wikidata

Кандидаты из поиска Wikipedia ищутся по Wikidata. Для каждого кандидата из Wikidata получается информация о гиперонимах или объектах более высокого класса. Например, Москва – это столица, город в России, то есть Россия – это объект более высокого класса, чем Москва. Все типы объектов сортируются по частоте и записываются в промежуточную таблицу.

Фильтрация классов

По идентификаторам классов из базы получаются их названия, которые дополняют таблицу классов. Полуавтоматически в базе отбираются те, которые относятся к географии (по запросам с указанием подстроки, где подстрока – это слово «страна», «город» и так далее). Более сложные случаи («община Нидерландов», где община как группа людей или как поселение, нельзя отметить все, где есть община или Нидерланды) с омонимией запрашиваются и размечаются вручную.

Выбор кандидатов и определение места

Из кандидатов, отобранных на этапе поиска в Wikipedia, выбираются те, которые присутствуют в Wikidata с одобренным классом. Первый по очереди кандидат с пометкой «география» считается итоговым правильным ответом.

Для выбранных объектов собирается информация о географических координатах через библиотеку wikidata для взаимодействия с API этого ресурса. В рамках этой работы целью является геолокация на уровне страны. Для этого используются материалы контурных карт (GADM), которые содержат описание многоугольников границ стран. С помощью библиотек для работы с геоданными в Python geopandas и shapely производится соединение датафрейма с координатами объектов и датафрейма с координатами стран. Таким образом каждой точке ставится в соответствие страна, куда попадают эти координаты. Приходится отметить, что такое решение не избавляет данные от уже указанной исторической проблемы: экспонатам приписывается страна происхождения на основе современных государственных границ, а не соответствующих историческому состоянию. Тем не менее, этот подход позволяет дать современному исследователю представление о географическом распределении происхождения музейных объектов, что согласуется с нашей исследовательской задачей.

Проверка единообразия

Для проверки консистентности полученного результата вводится дополнительная проверка: все распознанные уровни географического положения должны относиться к одной стране. Если это условие выполняется, то географическое происхождение экспоната считается успешно определенным.

Результаты

В настоящий момент сложно оценить качество алгоритма без масштабной ручной проверки. Среди проблем, которые потенциально могут влиять на качество данных можно указать омонимию названий («СССР, г. Брест», «Франция, Брест»). В будущем планируется проводить более сложную проверку и фильтрацию кандидатов с учетом информации об остальных уровнях, тем самым отсеивая ошибочные совпадающие варианты.

Всего экспонатов с указанием места: 8040362 (50 %)

Распознано мест: 7931705 (98.6 %)

4.3 Техники и материалы

Техники и материалы, использованные при создании экспоната, записываются в поле базы в виде списка. Насколько можно судить по данным, формат записи предполагает перечисление через запятую (некоторые записи отображаются как список в JSON-представлении. Однако такой вид записи соблюдается не везде: во многих случаях разделение запятыми отсутствует. Унификации препятствуют такие особенности, как:

- опечатки («лудение» вместо «лужение», «х/бээ»)
- разнообразие аббревиатур, непонятных неспециалисту («х/к», «х/м»)
- непоследовательно примененные знаки препинания (“[бумага]”)
- высокая детализация («бумага с водяным знаком в виде леандра»)
- свободный порядок слов в именных группах («фотопечать черно-белая глянцевая», «черно-белая глянцевая фотопечать», «глянцевая ч/б фотопечать», «фотопечать глянцевая ч/б»)

Унификация техник и материалов для нас состояла в нормализации текста.

Алгоритм нормализации можно описать следующим образом:

1) разделение составных частей описания техники и материалов, не разделенных запятой: «Бумага. Печать»

2) исключение не имеющей отношения к делу информации в скобках («многослойная живопись (1.012.8.0009)»)

3) нормализация орфографической вариативности: пробелы, дефисы/тире, слэши, необязательные квадратные скобки («скобление, склеивание», несколько пробелов), («[пластик]», «формовка\\ обжиг», «х\б», «х/б», «чернобелая печать», «черно – белая пе-чать»)

4) исключение детализированной информации (номера типов бумаги, значения пробы у металлов)

5) унификация по принципу «мешка слов»: слова внутри описания техники лемматизируются и сортируются по алфавиту («печать типографская», «типографская печать» таким образом приводятся к одному представлению).

6) нормализация отдельных известных сокращений (х/б, ч/б, ф/б и некоторые другие) по составленному полуавтоматически списку аббревиатур и приписанных вручную соответствий.

7) вариантам ставятся в соответствие наиболее частотное вхождение – это решает проблему того, что в лемматизированном виде трудно воспринимать названия. («печать типографский цветной»)

В дальнейшем необходимо будет разработать более сложную классификацию с опорой на онтологию (металл -> золото -> золото X пробы). Также представляется разумным применить автоматическую коррекцию орфографии, обученную на этом же наборе.

4.4 Авторы

Еще одним важным параметром описания музейного объекта является автор. Информация об авторе неоднородна как по музеям в целом, так и в описании коллекций отдельных музеев. Имя автора может быть написано разными способами: с инициалами, с полной версией имени, в разных алфавитах, содержать годы жизни или даже краткую биографию (см. таблицу 4). Для того чтобы достичь консистентности хотя бы для известных авторов, необходимо связать упомянутые в Госкатологе имена с идентификаторами Wikidata. Это позволяет группировать предметы по автору.

Для поиска имен в базах открытых связанных данных информация об авторе должна быть предобработана: имена отделены от лишней информации (годы жизни, биография) и разделены, если имен в поле несколько (это частый случай). Чтобы сократить число будущих запросов, одинаковые вхождения группируются и сортируются в порядке убывания частотности, чтобы с помощью первых N запросов покрыть наибольшую часть базы.

Ресурс, который мы использовали для поиска имен, – VIAF (Virtual International Authority

Таблица 4. Примеры написания имени автора

Количество вхождений	Запись в Госкаталоге
1392	Толстой Федор Петрович
224	Толстой Ф. П.
130	Толстой Фёдор Петрович
50	Толстой Ф.
40	Федор Толстой
40	Толстой Федор Петрович , скульптор
16	Толстой, Фёдор Петрович
8	Ф. Толстой
8	Ф. П. Толстой Резчик А. Лялин
8	Толстой, Фёдор Петрович, по барельефу
6	Толстой Ф. П. , художник
6	Толстой Ф. П. Императорский фарфоровый завод
3	Толстой, Федор Петрович (1783-1873)
3	Толстой, Федор Петрович

File). Он содержит в том числе ссылки на Wikidata, кроме того, с его помощью удобно фильтровать категорию личного имени и тем самым исключить географические названия. Однако, к сожалению, этот ресурс через некоторое время блокирует автоматические запросы. На данный момент получилось извлечь информацию только о части персон.

Работа с поиском происходит в такой последовательности:

1. Запрашивается выдача для текстовой строки, предположительно содержащей имя, в разделе VIAF «Принятые формы имен»

2. Извлекаются уникальные идентификаторы полученных по запросу имен

3. Запрашивается выдача для текстовой строки, предположительно содержащей имя, в разделе VIAF «Имена лиц»

4. Извлекаются уникальные идентификаторы полученных по запросу имен

5. Находится пересечение полученных списков

6. Если пересечение больше 10, то считается, что невозможно идентифицировать персону; если меньше, то варианты (ID VIAF) сохраняются в пользовательскую базу для дальнейшей фильтрации с помощью Wikidata

Например, по запросу «Горький» в принятых формах имен сервис выдает длинный список, состоящий в основном из произведений + имя писателя. В именах лиц находится имя писателя, а также объекты, которые считаются с ним связанными.

Принятые формы имен: «Горький, Максим, 1868-1936», «Горький и его эпоха. Материалы и исследования», «Горький, Максим, 1868-1936. | Фома Гордеев | English | (Bernstein : 2005)»...

Имена лиц: «Горький, Максим, 1868-1936», «Андреев, Леонид, 1871-1919», «Найденов, Сергей Александрович, 1869-1922»…

При пересечении остается «Горький, Максим, 1868-1936», для которого можно получить идентификатор Wlkidata.

5. Описательная статистика унифицированных данных

Экспонатам в базе приписана отнесенность к большим категориям. Рассмотрим распределение объектов каталога по этим категориям.

5.1 Категории экспонатов

Последние три категории встречают всего в нескольких десятков музеев. При этом в Госкаталоге есть синонимичные им более общие категории. Скорее всего, такие «окказиональные» категории будет лучше объединить с «большими»: «предметы этнографии» осмысленно было бы присоединить к «предметы прикладного искусства, быта и этнографии», к той же категории следовало бы присоединить и «предметы этнографии». «Документы, редкие книги», очевидно, следовало бы разделить между «Документы» и «Редкие книги». Затруднительно сказать, какие экспонаты попадают в категорию «прочее». Возможно, дальнейшая работа по унификации записей Госкаталога позволит выделить новые категории, которые бы лучше описывали занесенные в базу данные.

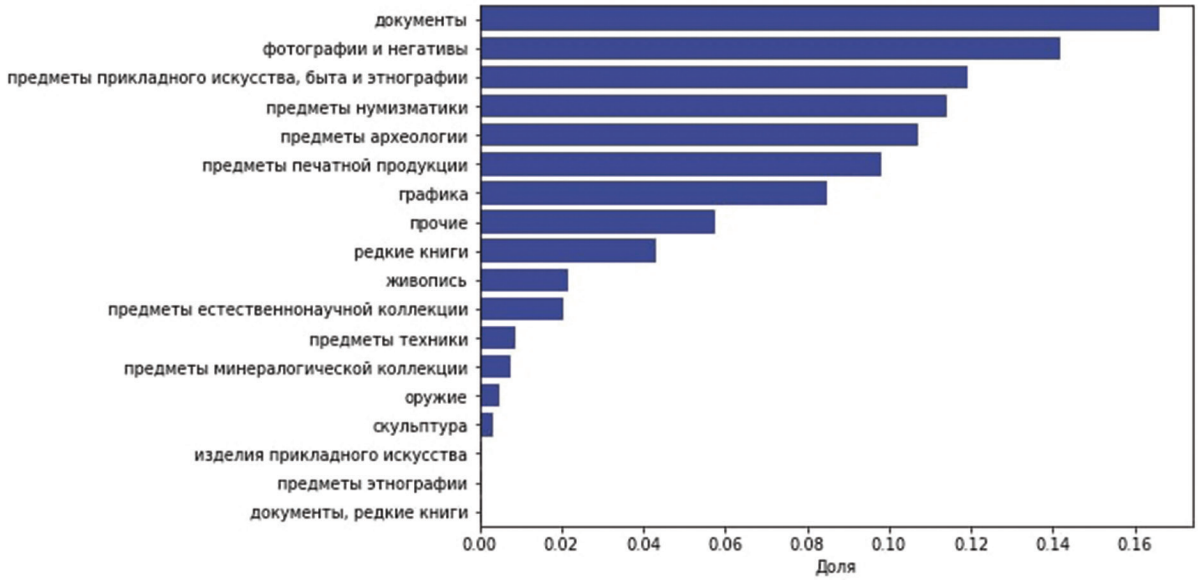


Рис. 1. Доли категорий экспонатов в коллекции Госкаталога

Таблица 5.

Количество экспонатов по категориям и число музеев, в которых эта категория представлена

Кол-во объектов/ музеев	Категория	Кол-во объектов/ музеев	Категория
2660516 (1149)	документы	348325 (1027)	живопись
2273742 (1112)	фотографии и негативы	329388 (222)	предметы естественнонаучной коллекции
1910158 (1373)	предметы прикладного искусства, быта и этнографии	141571 (783)	предметы техники
1828875 (1072)	предметы нумизматики	122250 (118)	предметы минералогической коллекции
1716150 (232)	предметы археологии	74703 (424)	оружие
1571897 (1132)	предметы печатной продукции	54123 (673)	скульптура
1362776 (820)	графика	8759 (47)	изделия прикладного искусства
925052 (1070)	прочие	8193 (35)	предметы этнографии
689915 (787)	редкие книги	6753 (32)	документы, редкие книги

5.2 Географическо-временное распределение экспонатов

Чтобы моделировать структуру коллекции, осмысленным представляется проанализировать, сколько экспонатов по периодам (векам, извлеченным в ходе унификации данных) и по странам присутствует в каталоге. Такое распределение может отражать значимость культурных, политических и экономических связей России в исторической перспективе.

Живопись

Поскольку история живописи хорошо известна, извлеченные из каталога данные позволяют проверить эффективность аналитического метода на этом материале. На графиках на рис. 3 можно увидеть отражение некоторых тенденций, свойственных нашему представлению об иерархии истории искусства. В какой-то мере эти графики напоминают аналогичные визуализации, используемые в культуромике, дисциплине, основанной на взаимодействии частотностей слов и культурно значимых событий:

- Авторитет итальянского Возрождения объясняет лидерство экспонатов из Италии вплоть до XVI века.
- XVII век считается золотым веком голландской живописи.
- эпоха романтизма, и эпоха импрессионизма связывается с подъемом изобразительного искусства Франции XVIII-XIX веков.
- Российско-персидские войны, закрепление геополитического влияния России в Персии в XIX веке, возросший интерес к Востоку, объясняющий пик Ирана на графике в XIX веке.

Графика

в XV-XVI веках лидирует немецкая графика, это так называемая «эпоха Дюрера».

- в Азии лидерами оказываются Китай и Япония, известные своими традициями графики. Пики для этих традиций отличаются, японский приходится на XIX век, а китайский на XX в.
- Большое число экспонатов французского происхождения может объясняться как ролью Франции как законодательницы художественной моды, так и тесным взаимодействием аристократий Франции и России в XIX веке.

Для остальных категорий распределения могут получиться значимыми к моменту, когда музейные данные в Госкаталоге будут представлены более полно.

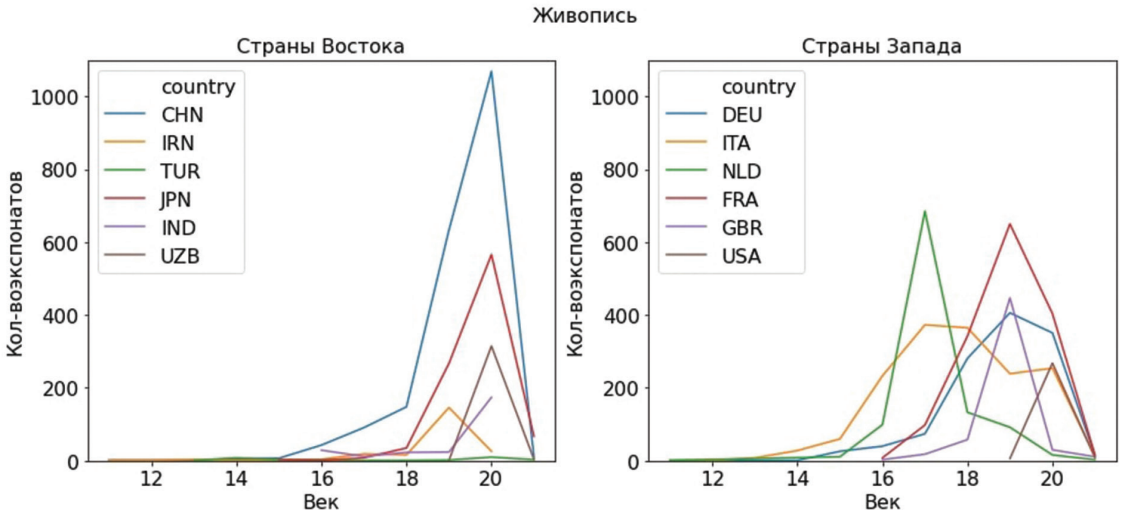


Рис 2. Распределение количества экспонатов категории «Живопись» по векам и странам. Коды стран: CHN – Китай, IRN – Иран, TUR – Турция, JPN – Япония, IND – Индия, UZB – Узбекистан; DEU – Германия, ITA – Италия, NLD – Нидерланды, FRA – Франция, GBR – Великобритания, USA – США

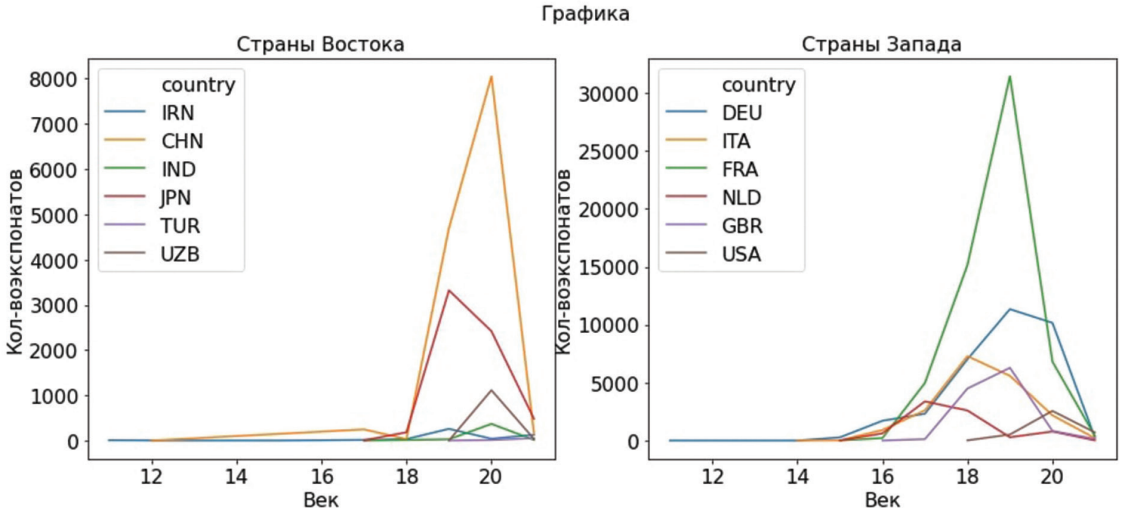


Рис 3. Распределение количества экспонатов категории «Графика» по векам и странам. Коды стран см. в подписи к рис. 2.

Разные категории представляют в целом общую картину перекосов данных внутри Госкаталога, хотя мы можем отметить и ряд окказиональных для каждой категории отличий. Среди восточных стран особенно выделяются Япония и Китай. Япония представлена в каталоге только начиная с XIX века, что объясняется политикой самоизоляции, которую страна проводила до этого времени. Китай имеет более длинную историю, выраженную в экспонатах российских музеев. В странах Запада (преимущественно Европы) можно заметить, что в ранний период (XIV-XVII вв) лидером оказывается Италия, в XVIII-XIX веках Франция, постепенно уступающая место Германии, захватившей в итоге лидерство почти во всех категориях. Стоит учитывать, однако, что на распределение экспонатов по тем или иным группам влияет целый ряд факторов:

- геополитическая обстановка
- художественная мода
- субъективные факторы, влияющие на формирование описания коллекции в каталоге

Под последний пунктом подразумевается, что мы не знаем, чем руководствуются музейные работники, отбирающие экспонаты для фиксации в Госкаталоге. Нужно учитывать, что этот процесс может быть растянут во времени, и коллекции пока ещё далеки от полноты описания. Таким образом, неочевидно, какие экспонаты (самые ценные? самые удобные для описания?) вносятся в базу в первую очередь. Возможно, дальнейшая работа с данными поможет ответить и на этот вопрос.

6. Выводы

Работа с базой показывает, что несмотря на особенности данных (заполнение на естественном языке, отсутствие шаблона или нестрогие правила), их можно успешно нормализовать и улучшить каталог.

На данный момент унифицированы даты, определены страны, где созданы экспонаты, частично обработаны техники и авторы. Результат работы с датами и странами показывает неплохое качество. Для работы с авторами и техниками результаты несколько хуже и требуют дополнительных интеллектуальных инструментов доводки.

Такая работа позволяет увидеть тенденции и особенности представления зарубежной культуры через представленность тех или иных стран и периодов в коллекции музеев. По статистике числа экспонатов можно увидеть, какие страны играли особую роль в культуре России и мира в определенные периоды. Статистика и подтверждает очевидные тенденции (например, роль итальянской, нидерландской и французской живописи).

References / Список литературы

Deuze M. Participation, remediation, bricolage: Considering principal components of a digital culture // The information society. – 2006. – Т. 22. – №. 2. – С. 63-75.

Said E. W. Invention, memory, and place //Critical inquiry. – 2000. – Т. 26. – №. 2. – С. 175-192.

Van Peer W., Hakemulder J., Zyngier S. Scientific methods for the humanities. – Amsterdam : Benjamins, 2012.

Правильная ссылка на статью

Глазунов Е. В., Орехов Б. В. Унификация данных музейного Госкаталога РФ // Сибирский антропологический журнал. – 2020. – Т. 4. – № 3 (09). – С. 154-168. Doi: 10.31804/2542-1816-2020-4-3-154-168