ОГЛАВЛЕНИЕ

ОТ СОСТАВИТЕЛЕЙ

Настоящий номер журнала «Электронные библиотеки» является второй частью тематического выпуска и включает статьи, подготовленные их авторами на основе материалов, представленных ими в 2020 году на XXII Всероссийской научной конференции «Научный сервис в сети Интернет».

Эта конференция была проведена с 21 по 25 сентября 2020 года и традиционно была посвящена направлениям и тенденциям использования интернет-технологий в современных научных исследованиях. Основная цель конференции — предоставить возможность для обсуждения, апробации и обмена мнениями о наиболее значимых результатах, полученных ведущими российскими учеными за последнее время в данной области деятельности. Организатором конференции был Институт прикладной математики им. М.В. Келдыша Российской академии наук. В связи со сложившейся эпидемической обстановкой конференция была проведена в режиме онлайн.

Первая часть тематического выпуска размещена в №1 журнала «Электронные библиотеки» за 2021 год, вторая часть – в настоящем номере.

М. М. Горбунов-Посадов, А. М. Елизаров

УДК 004.65 + 005 + 001.5

ИДЕНТИФИКАЦИЯ АВТОРОВ В РАМКАХ ПРЕДМЕТНОЙ ОБЛАСТИ В СЕМАНТИЧЕСКОЙ БИБЛИОТЕКЕ

О. М. Атаева[1, [0000-0003-0367-5575]], В. А. Серебряков[2, [0000-0003-1423-621X]], Н. П. Тучкова[3, [0000-0001-6518-5817]]

[1,2,3]*Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН, г. Москва*

[1]oli@ultimeta.ru, [2]serebr@ultimeta.ru, [3]natalia_tuchkova@mail.ru

Аннотация

Рассмотрены особенности задачи идентификации авторов и определения авторского вклада в публикации в цифровых библиографических коллекциях. Особенности проблемы недостаточной идентификации проявляются в повторах информации, двойниковании, наличии авторов с полностью совпадающими именами, самоцитировании, автоплагиате и собственно плагиате. Предлагается использовать информацию о публикациях, которая уже накоплена в цифровой библиотеке в виде связанных данных предметной области и множества данных тезауруса адресата, как автора и пользователя библиотеки. Эта информация содержит связи, благодаря которым для идентификации авторства можно использовать контексты ключевых слов, множества соавторов и ассоциативные связи терминов в словарях и тезаурусах. Важно, что рассматривается массив научных публикаций, поскольку они имеют сложившуюся традиционную структуру, что позволяет сравнивать фиксированные элементы текста (аннотации, ключевые слова, коды классификаторов и т. д.). Таким образом, даже при полном совпадении имен в публикациях можно ставить вопрос об авторстве, если в цифровой библиотеке публикации соответствуют различным предметным областям. Разрешение таких противоречий осуществляется путем оценки множества связей всех элементов вторичной информации о публикации. Результатом сравнения может быть добавление автора в некоторую предметную область, т. е. расширение тезауруса адресата и персонального тезауруса автора, или появление в библиотеке полных тезок, но из разных областей знаний. Показано, что современные средства анализа данных

позволяют оценить вклад автора в публикацию, несмотря на то, что конечно, реальный вклад в научное исследование может оценить только научное сообщество.

Ключевые слова*: сравнение научных текстов, семантический поиск, тезаурус для онтологии знаний*, *информационный запрос с помощью тезауруса, семантические библиотеки, способы идентификации авторов, тезаурус адресата, вторичная информация, частотный словарь индивидуума***,** *LibMeta*.

ВВЕДЕНИЕ

Проблемы определения того, кто заслуживает быть автором научной статьи и каков его вклад в коллективную публикацию, если в цифровой коллекции нет достоверной информации, разрешаются различными способами. В основном выполняются сопоставление близких по тематике статей и опрос зарегистрированных авторов, как в ResearchGate. С вопросами, связанными с идентификацией авторов в библиографических системах, сталкиваются практически все цифровые ресурсы, известные на сегодняшний день. При обновлении информации могут появиться «спорный» автор, полный тезка, «старый» автор с другой транскрипцией в написании фамилии и т. д. Всем известны трудности собственной идентификации даже в таких авторитетных базах данных, как WoS и Scopus, когда несмотря на все выставленные фильтры, получаем в результате поиска список из «смеси» своих и чужих работ, что отражено, например, в публикации [1]. Нередко приходится вручную формировать необходимый список, несмотря на существующий в этих системах (как и во многих других) механизм автоматического формирования авторского указателя. Исключение составляют публикации и издания, в которых изначально требуется задать ORCID автора. Собственные идентификаторы ввели также elibrary (SPIN-код автора), система ИСТИНА (IstinaResearcherID, IRID), Scopus (Scopus Author ID), Web of Science ResearcherID, Google Scholar Citation ID. Чем больше индексов указывает автор при регистрации в этих системах и статьях при передаче их издательствам, тем точнее он идентифицируется, естественно. Некоторые издательства делают обязательными ссылки на индексы авторов соответствующих баз данных, с которыми эти издательства сотрудничают. Тот факт, что идентификаторы авторов сопровождают публикации, говорит о том, что другие

способы, несмотря на принятые правила идентификации, оказываются недостаточно надежными.

Существует ряд требований к статьям и авторам в отдельных специфических предметных областях, и они были утверждены, например, для авторства в медицинских исследованиях, но стали впоследствии общепринятыми. Автор – это тот, кто участвует в развитии идеи, сборе и анализе данных, написании работы, внесении в текст актуальных и идеологически оправданных изменений.

Тем не менее, для определения вклада автора в коллективные исследования этих средств недостаточно, на что указывается, например, в работе [2]. Более того, в цифровой век в некоторых научных сообществах существуют варианты: обсуждение коллегами вклада авторов в исследования; предоставление издательствам права высказывать мнение об авторстве на основе накопленной информации. Это нарушает традиционные нормы, принятые ранее [3].

Изменился уровень достоверности, прозрачности и документирования данных об авторах. Таким образом, проблема авторства ставится шире и не ограничивается вторичной информацией при индексации в базах данных. Эта проблема включает человеческий фактор, опрос экспертов, редакторов и соавторов. В целом отмечается тенденция увеличения числа соавторов за последние 30 лет [4], хотя для отечественных научных работников это ведет к известным проблемам в отчетности перед фондами и министерствами.

В настоящей работе рассматриваются варианты использования данных, которые имеются в арсенале современных информационных технологий для индексации публикаций, авторов и их вклада в коллективные работы в цифровых библиографических системах.

1. О СРЕДСТВАХ ИДЕНТИФИКАЦИИ АВТОРОВ

1.1. Множества данных для идентификации авторов

Структура научной публикации – это особенность научных статей, вполне устоявшаяся для многих отечественных и международных журналов. Строгость, которой предлагается придерживаться авторам в соответствии с инструкцией от издателей, продиктована в какой-то мере процессом оцифровки публикаций для последующей их индексации в библиографических базах данных. В 1070-х годах появилось семейство стандартов для машиночитаемой каталогизации (*Machine-*

Readable Cataloging, MARC) [5] с дальнейшей разработкой стандарта ISO 2709 (ГОСТ 7.14-84 (СТ СЭВ 4269-83) СИБИД и ГОСТ 7.14-98 СИБИД). Эти стандарты первоначально были предложены Библиотекой конгресса США в качестве форматов межбиблиотечного обмена библиографическими данными, а позднее адаптировались для национальных библиотек и стали в той или иной форме использоваться во всех англоязычных библиотечных системах. Естественным образом стандартные поля библиографических записей для машиночитаемой каталогизации стали компонентами и фиксированными позициями в структуре научных статей.

Таким образом, был сформирован список обязательных полей вторичной информации о документе «научная статья»: автор(ы), аффиляция автора(ов), название, ключевые слова, классификаторы (MSC, UDC и/или специализированные), выходные данные (издательство, страницы, год). В дальнейшем добавились аннотация, список цитируемой литературы и идентификаторы, такие как ORCID и др. Все эти поля используются для индексирования публикаций и могут участвовать в качестве поисковых при формировании запроса и идентификации авторов.

Трудность возникает, если этой информации недостаточно или ее нет в полном объеме в базе данных или у пользователя. Уточнение осуществляется благодаря экспертным знаниям или за счет семантических связей, которые могут быть реализованы в виде подсказок из базы данных.

Тело публикации, как правило, недоступно для поиска, даже если публикация находится в открытом доступе, но доступно издателям для предварительной лексической, синтагматической, парадигматической, семантической обработки при размещении в библиографических базах данных.

1.2. Набор данных тезауруса адресата

Понятие «адресата в информационной среде», сформулированное для удобства определения пользователей и авторов из баз данных, подразумевает персону – участника информационного процесса, поиска и обмена информацией. Термин «тезаурус адресата (индивидуума)» (ТА) введен в информатику Ю.А. Шрейдером [6] для представления предметной области (ПрО) автора на основе понятийного запаса знаний автора. Термин связан также с представлением «знаний» в информационной системе как «структурированной информации» [7].

Для более подробного знакомства с использованием тезаурусов в поисковых процессах и извлечения знаний можно обратиться к работе [8]. В дальнейшем проявилась важность этого представления, как основы для описания онтологии адресата (ОА) в современных базах данных [9].

Состав данных (информации) тезауруса адресата зависит от понятийного запаса индивидуума. Можно остановиться на следующем наборе данных: частотный словарь индивидуума; варианты сочетаний терминов; контексты частотных терминов; специальные обозначения и формулы; списки цитируемой литературы; списки цитирующих авторов; список публикаций с перекрестными ссылками. Если в информационной системе достаточно данных и публикаций по некоторой предметной области, то на основе множества данных о тезаурусе адресата и метрического анализа можно построить *словарь-тезаурус предметной области автора*. Далее, сравнивая предметные тезаурусы авторов, можно более точно их идентифицировать, а также устанавливать принадлежность текста некоторому автору и его вклад в исследования.

1.3. Инструменты сравнения текстов для идентификации авторов

Рассматриваются методы сравнения текстов для установления авторства, такие как частотные алгоритмы [10], контекстное сравнение [11], тематическая кластеризация и алгоритмы глубокого анализа текстов, связанные с методами машинного обучения [12], [13].

Используя эту совокупность методов, можно сформировать технологию обработки информации для *вновь поступающих данных* в информационную библиографическую систему.

Первый этап предварительной обработки (препроцесса) публикаций для *каждого автора* включает:

- частотную обработку текстов для получения списка терминов с их весом (частотой использования);
- составление списка соавторов;
- формирование множества контекстов для терминов.

В результате накапливаются следующие данные (параметры) *автора*: список (словарь) терминов, ранг (вес) терминов, словоформы терминов, относительная частота терминов (по отношению к другим терминам), абсолютная частота

терминов, конкорданс словарь (словарь с контекстами), рис. 1. На этом этапе также возможно выделить список уникальных терминов, обозначений, формул и других особенностей текста, характерных для некоторых авторов и предметных областей.

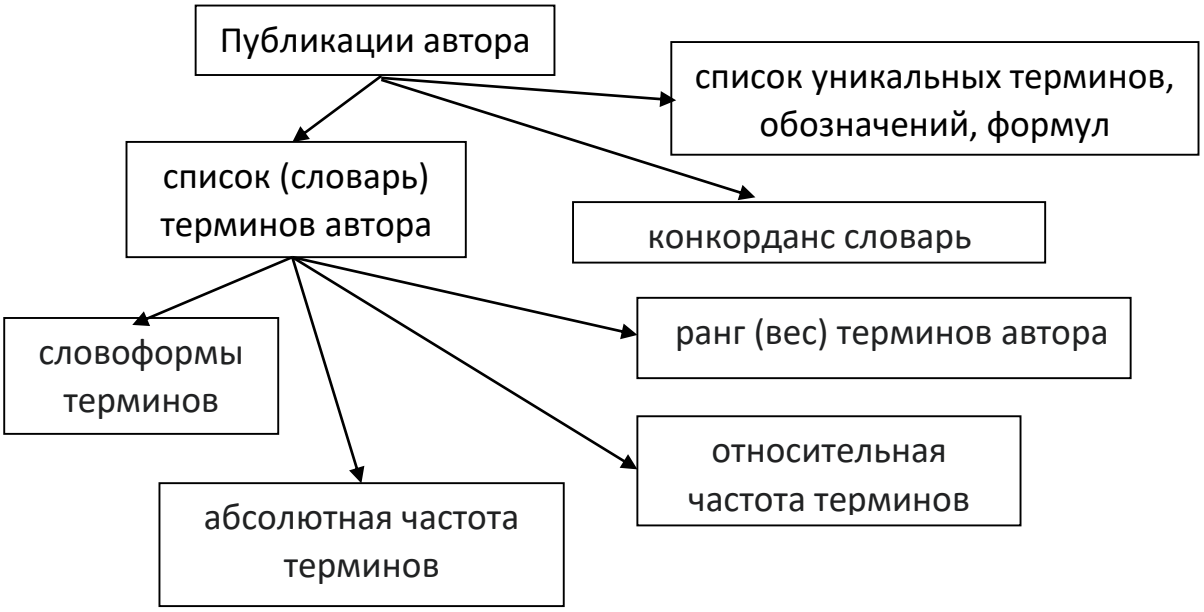


Рис. 1. Схема предварительной обработки публикаций автора.

Второй этап заключается в процедуре сравнения авторов по имеющимся (накопленным) параметрам. Выявляются пересечения множеств терминов, контекстов и уникальных терминов, обозначений и т. д.

После сравнения и выявления множества публикаций, принадлежащих определенному автору, составляются авторский указатель и указатель цитируемых публикаций. При этом можно варьировать строгость принадлежности «спорных» публикаций тому или другому автору, учитывая степень совпадений выявленных параметров (в %, например).

На этом предварительная обработка *вновь поступающих данных об авторе* может быть закончена.

Все это множество связанной полученной информации можно считать тезаурусом адресата.

Замечание 1. Если в систему предполагается загрузить *серию публикаций одного автора* (или авторского коллектива), то можно на предварительном этапе обработки составить тезаурус адресата (адресатов).

Замечание 2. Если поступила единичная работа, то предварительная обработка (по схеме рис. 1) используется для включения в имеющийся авторский указатель или при отсутствии совпадений и спорных свойств публикации (варианты фамилий и других вторичных документов) хранится в статусе подтверждения, но участвует в дальнейшей предметной семантической обработке. Подтверждение можно делать автоматически, если в системе накопится дополнительная информация об авторе или по запросу к автору.

Для дальнейшей семантической обработки публикаций необходимо использовать словари (тезаурусы) профессиональных терминов из предметных областей (например, математических).

Публикации необходимо проиндексировать в соответствии с предметной и тематической направленностью, определяя принадлежность терминов публикаций словарям (тезаурусам) предметных областей. Таким образом можно зафиксировать связи тезауруса адресата (автора) с предметными областями. Эти связи представляют в дальнейшем дополнительные *признаки для предметной идентификации автора*.

Таким образом, публикации, связанные семантически в онтологиях, в результате препроцессорной обработки будут иметь еще ряд признаков идентификации авторов.

2. ПРИМЕРЫ НА НАБОРАХ ДАННЫХ

На примере некоторого множества работ по разделам высшей математики можно рассмотреть варианты идентификации авторов публикаций со схожими наборами вторичных документов.

Для обработки текста используется свободная библиотека для высокопроизводительного полнотекстового поиска Apache Lucene, реализованная на языке Java.

2.1. Установление авторства

Для выделения значимых выражений документа использовался расчет меры tf-idf для терминов документа, извлеченных из индекса, с учетом морфологии [13]. На первом этапе рассматривались только существительные и термины, которые были идентифицированы как имена собственные.

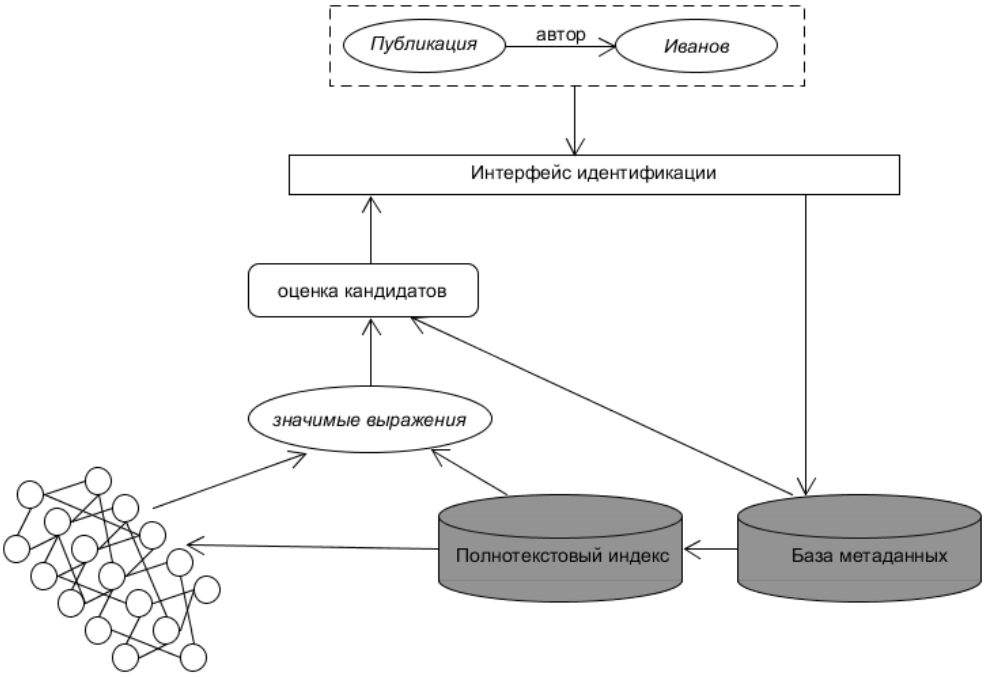


Рис. 2. Общая схема работы с терминами и авторами.

Далее исключались термины, для которых мера tf-idf была меньше порогового значения. Составление комбинаций из двух и трех слов выполнялось на основе использования контекста выделенных слов и правил, учитывающих морфологию. Под контекстом понимаются N слов, находящихся в тексте перед словом, для которого строится вектор, и N слов, находящихся после этого слова. Для выделения контекста используется неглубокая нейросетевая модель word2vec [14]–[16] в режиме «skip-grams». На рис. 2 представлена общая схема работы.

В качестве примера далее на рис. 3 отражен этап формирования тезаурусов предметных областей отдельных авторов, на основе которых можно рассуждать об их (авторов) идентичности.

Из примера видно, что были получены работы авторов с неполным набором вторичной информации. Применение описанного алгоритма позволяет выявить термины, связи и пересечения подмножеств терминов с учетом их контекстов.

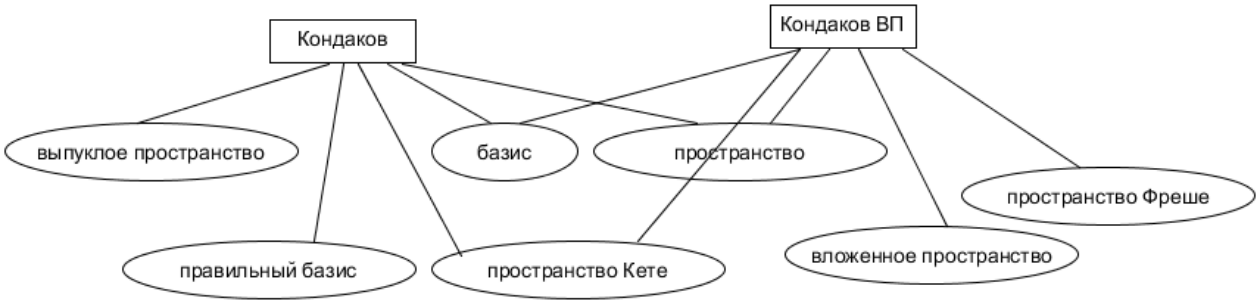


Рис. 3. Общая схема сравнения авторов.

Используем далее дополнительно связи терминов из энциклопедии, классификаторов УДК, MSC и других работ из области аналитических пространств, такие, как представлены на рис. 4.

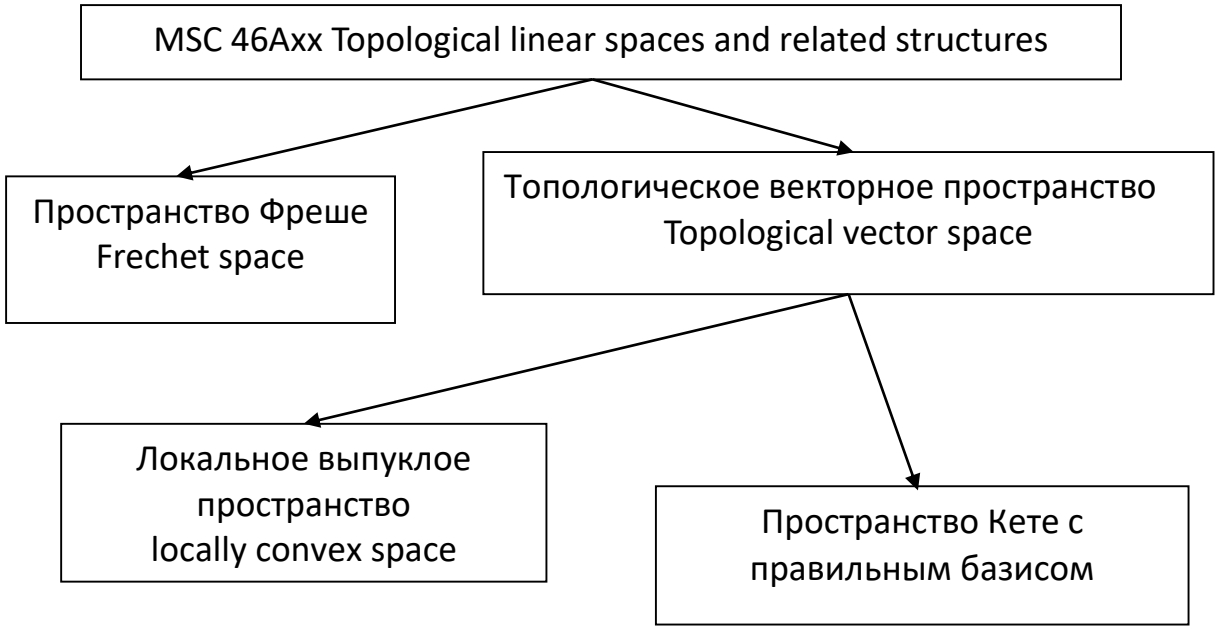


Рис. 4. Связи выявленных терминов авторов.

Было обработано около 5000 авторов публикаций. Отдельно проводится работа по обработке формул и включения их в тезаурус автора. Используется алгоритм сравнения формул на основе векторной модели. Алгоритм условно делится

на две части: первичный отбор формул-кандидатов и последующее их упорядочение по схожести. Описание этого алгоритма выходит за рамки данной статьи.

2.2. Учет авторского вклада

Для учета авторского вклада в публикацию требуется исследовать историю работ автора и его принадлежность научным школам, а также исследования автора в предметных областях. Это имеет особое значение, поскольку соавторство стало носить коммерческий характер и стали возможны платные публикации, «старшее авторство» и цитирование [17].

Совокупность «исторических» данных об авторе и публикации формируется на основе тезауруса адресата следующим образом. Собирается и хранится история публикаций: соавторы, перекрестные ссылки, ключевые слова, внутренние системные индикаторы принадлежности публикаций предметным областям (LibMeta). Практически все современные библиографические коллекции собирают и демонстрируют перечисленные данные. Развитие информационных технологий позволяет привлечь различные методы анализа для установления авторства публикаций. Надо отметить, что для художественных произведений типа «карманных» детективов такая экспертиза проводится давно, поскольку этот процесс изначально построен на коммерческой основе, и необходим учет вклада каждого участника. Для научного сообщества требуется избегать такого подхода, так как это неизбежно ведет к принижению значения исследовательской работы.

Выбраны признаки, по которым распределяются публикации, это история вопроса, новизна, количество публикаций на смежные темы, множества соавторов, экспертное мнение, выраженное в процессе дискуссий и рецензирования.

История публикации. Структура научной статьи предполагает наличие вступления, в котором перечисляются предыдущие исследования. Анализируя этот текст, можно составить списки исследователей и соответствующих библиографических ссылок по выбранной теме, пример – рис. 5.

Далее нужно выбрать пересечения внутри этих множеств и выявить «главных» авторов и их соавторов. Для соавторов выявить частотные характеристики и принадлежность предметной области. Таким образом, получить «карту» публикаций по теме, где будут области пересечения авторских коллективов, где пересекаются {k1, k2, …, kN} авторов (k1>k2> …>kN). В эти области включаются и авторы

из списков цитирования. Отдельно стоящие авторы (A, B, C, …) могут принадлежать множествам «приглашенных» к участию в публикациях, и тогда их роль оценивается экспертами из научного сообщества. Это могут быть авторы публикаций без соавторов, работающие в данной предметной области, и тогда, естественно, их вклад в работу не оспаривается.

Рис. 5. Общая схема сравнения авторов.

Множество авторов k1, которое больше других, может претендовать на множество ведущих ученых, руководителей научных школ и исследовательских проектов (грантов и пр.).

Оценка по ключевым словам. Пересечение ключевых слов в тезаурусах авторов свидетельствует о близости исследований.

Новизна. Анализ коллективов, составляющих множества {k1, k2, …, kN}, позволяет выявить «новых» членов авторского коллектива, за какой-то период времени, «новые» ключевые слова за этот же период времени. Поскольку благодаря тезаурусу адресата можно выяснить, к какому автору относятся «новые» ключевые слова, то можно сделать вывод о том, благодаря кому появился «новый» вклад в публикации и предметную область.

На основе данных тезауруса адресата в системе LibMeta введена метрика оценки авторского участия в публикации по математическим предметным областям. Вычисляются следующие множества:

- ядро ключевых концептов предметной области (Concept Kernel) – {СК = $К_1 \cap К_2 \cap К_3$ }, где $К_1$ – тезаурус ОДУ, К2 – словарь спец. функции и К3 – математической энциклопедии:

$$|KK|=|K_1|+|K_2|+|K_3|, |K_1|=184, |K_2|=151, |K_3|=6263, |KK|=6\,598,$$

- ядро ключевых слов информационных объектов для разных типов ресурсов ресурсов предметной области (Keyword Kernel) – {KK}, |KK|=6810,

- ядро авторских коллективов по годам (Kernel of Copyright Teams) – {КСТ}.

Рассмотрим для примера 2015 год для публикаций, затрагивающих *обыкновенные дифференциальные уравнения Бернулли*[1].

Получаем: $|KK_{2015}|=754$, $КСТ_{2015}$={*'Лазарев'*, *'Неустроева'*, *'Шишкина'*, *'Бочкарев'*, *'Лекомцев'*, *'Сенин'*, *'Янковский'*, *'Кольцун'*}, ядро библиографических ссылок (Bibliographic Reference Kernel) – {BRK} для этих авторов представлено 34 ссылками, |BRK|=34.

Далее оценивается пересечение данных из ТА автора: ключевых слов {KWA}, $|KWA_{Лазарев}|=14$, $|KWA_{Янковский}|=79$, соавторов {CA} $|CA_{2015}|=163$, библиографических списков {RL} $|RL_{Лазарев}|=3$, $|RL_{Янковский}|=16$, с множествами $\{KK_{2015}\}$, $\{КСТ_{2015}\}$, $\{BRK_{2015}\}$ к общим характеристикам предметной области:

$$\{KWA\} \cap \{KK\}, \{CA\} \cap КСТ\}, \{RL\} \cap \{BRK\}.$$

На основании этого вводятся оценки вклада автора в предметную область $KWA_{Лазарев}/KK_{2015}=14/754$, «средний» вклад автора в предметную область в этом году $CA_{2015}/КСТ_{2015}=163/8$, «средний» вклад $|RL_{Лазарев}|/|BRK|=3/34$, $|RL_{Лазарев}|/|BRK|=16/34$.

Эти оценки показывают вклад автора в предметную область и конкретные исследования (публикации) «во времени». Подчеркнем, что эти оценки не отражают картину реального мира, но справедливы для характеристики того множе-

[1] http://libmeta.ru/concept/showRelatedValues/404?attribute=119

ства объектов, которые загружены в систему. Авторы, у которых наибольший процент «пересечений» с онтологией ПрО, могут считаться «ключевыми» исследователями в предметной области.

Рассмотрим матрицу (Таблица 1) признаков публикации по новизне, где критерий – это новые ключевые слова.

Таблица 1: Соответствие *Ключевых слов* предметной области публикациям и авторам

	публикации (art)	сравнение {art} и {artПрО}	ПрО цифровой библиотеки	содержит art%
	авторы (au)	сравнение {au} и {auПрО}		содержит au%
Ключевые слова	тезаурусы автора (auths)	сравнение {auths} и {thsПрО}		содержит auths%
	UDC (udc)	сравнение {udc} и {udcПрО}		содержит udc%
	MSC (msc)	сравнение {msc} и {mscПрО}		содержит msc%
	Формулы (form)	сравнение {form} и {formПрО}		содержит form%

Таблица 1 многомерная и содержит наибольшее количество возможных связей ключевых слов. В ней присутствует связь *ключевых слов* с авторским предметным тезаурусом (если он есть) и тезаурусом ПрО, который заложен в основу онтологии ПрО в семантической библиотеке. На основе сравнения множеств из столбцов получаем значения (например, в процентном отношении) в последнем столбце и принимаем решение о принадлежности ключевых слов ПрО семантической библиотеке или о новом множестве для этой библиотеки и ПрО, т. е. можно принять решение о том, насколько новая публикация соответствует ПрО.

Замечание 3: В данном исследовании не дается никакой оценки обоснования исследований авторов и качества научных работ.

Замечание 4: Все оценки делаются только на основе публикаций, вторичной информации или полных текстов (если они доступны) и авторских методов, отслеживания связей в цифровой библиотеке.

Замечание 5: Реальный вклад автора в публикацию и исследования может оценить только научное сообщество. В цифровой библиотеке можно установить количество и тип связей по выбранным признакам и на основе массива данных, который есть в библиотеке. Этот анализ дает картину вклада публикации и рейтинг автора в масштабах имеющихся данных, но не качества публикации и знаний автора в целом.

Замечание 6: В библиотеке LibMeta используется технология создания предметного авторского тезауруса, и на его основе можно получить представление о тезаурусе адресата как участника обмена информацией в информационной среде. Эта технология позволяет рассматривать *значение и вклад* публикаций автора применительно к различным предметным областям, которые составляют пересечение множеств в рамках предметного авторского тезауруса.

В работе представлена идеальная схема оценки роли автора и установления авторства, конечно, в ней есть спорные факторы, но схема может быть использована как первое приближение, если авторство статьи вызывает сомнения по причине неточности вторичных данных в цифровой библиотеке.

В реальности провести границу между претендентами на авторство публикации может быть непросто, что иногда является предметом спора научных школ. Известны случаи, когда идея и ее реализация в исследованиях принадлежат различным людям, которые могут знать или не знать о работах друг друга. Здесь поднимаются вопросы плагиата и приоритетов в науке. Пример тому являет история разногласий Ньютона и Лейбница по вопросу вклада каждого в развитие математического анализа [18]. Именно в цифровых библиотеках можно учесть, если не все, то многие признаки авторства, что показано на примерах математических статей в LibMeta.

ЗАКЛЮЧЕНИЕ

Предложена технология предварительной обработки публикаций для дальнейшего размещений в семантической библиотеке. Использование данных теза-

уруса адресата позволяет накапливать структурированную информацию об авторах и публикациях, что способствует на предварительном этапе идентифицировать авторов и оценить их вклад в исследования.

Использование персональной среды для научного исследования на базе индивидуальных библиографических коллекций и результатов, собранных автором в процессе исследований, позволяет рассматривать задачи идентификации и определения авторского вклада как часть функционирования семантической библиотеки.

Благодарности

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект № 20-07-00324, и в рамках темы Министерства науки и высшего образования РФ «Математические методы анализа данных и прогнозирования».

СПИСОК ЛИТЕРАТУРЫ

1. *Krämer T., Momeni F., Mayr P.* Coverage of Author Identifiers in Web of Science and Scopus. – arXiv preprint arXiv:1703.01319, 2017 – arxiv.org.

Clement T.P. Authorship Matrix: A Rational Approach to Quantify Individual Contributions and Responsibilities in Multi-Author Scientific Articles // Science and Engineering Ethics. 2014. V. 20. P. 345–361. URL: https://doi.org/10.1007/s11948-013-9454-3.

2. *Frische S.* It is time for full disclosure of author contributions// Nature. 2012. P. 489.

URL: http://www.nature.com/news/it-is-time-for-full-disclosure-of-author-contributions-1.11475.3.

3. *Cozzarelli N.R.* Responsible authorship of papers in PNAS // Proceedings of the National Academy of Sciences of the United States of America. 2004. V. 101, No. 29. P. 10495.

4. MARC 21 Formats. URL: http://www.loc.gov/marc/marcdocz.html.

5. *Шрейдер Ю.А.* Тезаурусы в информатике и теоретической семантике // Научно-техническая информация. Сер. 2. 1971. № 3. С. 21–24.

6. *Гаврилова Т.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем. СПб.: Питер, 2000. 384 с.

7. *Лукашевич Н.В.* Тезаурусы в задачах информационного поиска. М.: Изд-во МГУ, 2011. 495 с.

8. *Муромский А.А., Тучкова Н.П.* Об онтологии адресата в математической предметной области // Электронные библиотеки. 2018. Т. 21, № 6. С. 506–533.

9. *Борисов Л.А., Орлов Ю.Н., Осминин К.П.* Идентификация автора текста по распределению частот буквосочетаний // Препринты ИПМ им. М.В. Келдыша. 2013. № 27. 26 с.
URL: http://library.keldysh.ru/preprint.asp?id=2013-27.

10. TextSTAT - Simples Text Analyse Tool. URL: http://neon.niederlandistik.fu-berlin.de/textstat/.

11. *Mohsen A.M., El-Makky N.M., Ghanem N.* Author Identification Using Deep Learning, 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, 2016. P. 898–903.
URL: https://doi.org/10.1109/ICMLA.2016.0161.

12. *Маннинг К.Д., Рагхаван П., Шютце Х.* Введение в информационный поиск. Cambridge University Press, 2018. 482 p.

13. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR, 2013.

14. *Mikolov T., Yih W.T., Zweig C.* Linguistic Regularities in Continuous Space Word Representations // Proceedings of NAACL HLT, 2013.

15. *Le Q., Mikolov T.* Distributed Representations of Sentences and Documents // International Conference on Machine Learning, 2014. P. 1188–1196.

16. *Strange K.* Authorship: Why not just toss a coin? // American Journal of Physiology-Cell Physiology. 2008. V. 295, No. 3. P. 567–575.
URL: https://doi.org/10.1152/ajpcell.00208.2008.

17. *Meli D.B.* Equivalence and Priority: Newton versus Leibniz: Including Leibniz's Unpublished Manuscripts on the Principia. Clarendon Press, 1993. P. 318.

AUTHORS IDENTIFICATION WITHIN THE SUBJECT AREA IN THE SEMANTIC LIBRARY

O. M. Ataeva[1, [0000-0003-0367-5575]], V. A. Serebriakov[2, [0000-0003-1423-621X]], N. P. Tuchkova[3, [0000-0001-6518-5817]]

[1,2,3]*Dorodnicyn Computing Centre FRC CSC RAS, Moscow*

[1]oli@ultimeta.ru, [2]serebr@ultimeta.ru, [3]natalia_tuchkova@mail.ru

Abstract

The peculiarities of the task of authors identifying and determining author's contribution to publications in digital bibliographic codes are considered. The features of the problem of insufficient identification are manifested in the repetition of information, doubling, the presence of authors with completely coincidental names, self-quotation, autoplagiate and plagiarism itself. It is proposed to use publication information that has already been accumulated in the digital library in the form of related object area data and a variety of target thesaurus data, as the author and user of the library. This information contains links whereby keyword contexts, multiple co-authors, and term associations in dictionaries and thesauruses can be used to identify authorship. It is important that an array of scientific publications is considered, since they have an established traditional structure, which allows comparing fixed text elements (annotations, keywords, classifier codes, etc.). Thus, even if the names in the publications are fully matched, the question of authorship can be raised if the publications in the digital library correspond to different subject areas. Resolution of such contradictions is accomplished by evaluating a plurality of links of all elements of secondary publication information. The result of the comparison could be the addition of the author to a specific area, i.e. the extension of the addressee's thesaurus and the author's personal thesaurus, or the appearance of full namesakes in the library, but from different areas of knowledge. It has been shown that modern data analysis tools allow you to evaluate the author's contribution to publication, despite the fact that of course, only the scientific community can evaluate the real contribution to scientific research.

Keywords: *comparison of scientific texts, semantic search, thesaurus for the ontology of knowledge, information query using the thesaurus, methods of authors identification*, *addressee thesaurus, secondary information, individual frequency dictionary, LibMeta*.

REFERENCES

1. *Krämer T., Momeni F., Mayr P.* Coverage of Author Identifiers in Web of Science and Scopus. – arXiv preprint arXiv:1703.01319, 2017 – arxiv.org.

2. *Frische S.* It is time for full disclosure of author contributions// Nature. 2012. P. 489.
URL: http://www.nature.com/news/it-is-time-for-full-disclosure-of-author-contributions-1.11475.3.

3. *Cozzarelli N.R.* Responsible authorship of papers in PNAS // Proceedings of the National Academy of Sciences of the United States of America. 2004. V. 101, No. 29. P. 10495.

4. MARC 21 Formats. URL: http://www.loc.gov/marc/marcdocz.html.

5. *Shrejder Yu.A.* Tezaurusy v informatike i teoreticheskoj semantike // Nauchno-tekhnicheskaya informaciya. Ser. 2. 1971. № Z. S. 21–24.

6. *Gavrilova T.A., Horoshevskij V.F.* Bazy znanij intellektual'nyh si-stem. SPb.: Piter, 2000. 384 s.

7. *Lukashevich N.V.* Tezaurusy v zadachah informacionnogo poiska. M.: Izd-vo MGU, 2011. 495 s.

8. *Muromskij A.A., Tuchkova N.P.* About ontology of the addressee in mathematical subject domain // Russian Digital Library Journal. 2018. V. 21, № 6. P. 506–533.

9. *Borisov L.A., Orlov Yu.N., Osminin K.P.* Identifikaciya avtora teksta po raspredeleniyu chastot bukvosochetanij // Preprinty IPM im. M.V. Keldysha. 2013. № 27. 26 s. URL: http://library.keldysh.ru/preprint.asp?id=2013-27.

10. TextSTAT - Simples Text Analyse Tool. URL: http://neon.niederlandistik.fu-berlin.de/textstat/.

11. *Mohsen A.M., El-Makky N.M., Ghanem N.* Author Identification Using Deep Learning, 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, 2016. P. 898–903.
URL: https://doi.org/10.1109/ICMLA.2016.0161.

12. *Manning K. D., Raghavan P., Schutze H.* Introduction to Information Retrieval. Cambridge University Press, 2018. 482 p.

13. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR, 2013.

14. *Mikolov T., Yih W.T., Zweig C.* Linguistic Regularities in Continuous Space Word Representations // Proceedings of NAACL HLT, 2013.

15. *Le Q., Mikolov T.* Distributed Representations of Sentences and Documents // International Conference on Machine Learning, 2014. P. 1188–1196.

16. *Strange K.* Authorship: Why not just toss a coin? // American Journal of Physiology-Cell Physiology. 2008. V. 295, No. 3. P. 567–575.
URL: https://doi.org/10.1152/ajpcell.00208.2008.

17. *Meli D.B.* Equivalence and Priority: Newton versus Leibniz: Including Leibniz's Unpublished Manuscripts on the Principia. Clarendon Press, 1993. 318 p.

СВЕДЕНИЯ ОБ АВТОРАХ

АТАЕВА Ольга Муратовна – научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, кандидат техн. наук, специалист в области системного программирования и баз данных.

Olga Muratovna ATAEVA – researcher of the of Dorodnicyn computing center FRC SCS RAS, PhD, expert in the field of system programming and databases.

email: oli@ultimeta.ru

СЕРЕБРЯКОВ Владимир Алексеевич – специалист в области теории формальных языков и её приложений, доктор физ.-мат. наук, профессор, зав. отделом Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН. Руководитель и участник разработки ряда известных программных проектов, в частности, ИСИР и ИСИР РАН, «Научный портал РАН».

Vladimir Alekseevich SEREBRIAKOV – expert in the field of theory of formal languages and its applications, doctor of sciences, professor, head of Dorodnicyn computing center FRC SCS RAS department. Head and participant in the development of a number of well-known program projects, in particular, ISIR and ISIR RAS, Scientific portal RAS.

email: serebr@ultimeta.ru

ТУЧКОВА Наталия Павловна – старший научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, кандидат физ.-матем. наук, окончила ВМиК МГУ им. М.В. Ломоносова. Специалист в области алгоритмических языков и информационных технологий.

Natalia Pavlovna TUCHKOVA – senior researcher of Dorodnicyn computing center FRC SCS RAS, PhD in physics with a math degree, graduated from CS Faculty of Lomonosov MSU. The expert in the field of algorithmic languages and information technologies.

email: natalia_tuchkova@mail.ru

Материал поступил в редакцию 25 ноября 2020 года

УДК 013:004.65

ИНФОРМАЦИОННАЯ СИСТЕМА РЕГИСТРАЦИИ РЕЗУЛЬТАТОВ ИНТЕЛЛЕКТУАЛЬНОЙ ДЕЯТЕЛЬНОСТИ СОТРУДНИКОВ НАУЧНОГО УЧРЕЖДЕНИЯ

С. А. Власова[1, [0000-0003-1533-5850]], Н. Е. Каленов[2, [0000-0001-5269-0988]]

[1-2]Межведомственный суперкомпьютерный центр (МСЦ) РАН – филиал ФГУ ФНЦ Научно-исследовательский институт системных исследований (НИИСИ) РАН

[1]svlasova@jscc.ru, [2]nkalenov@jscc.ru

Аннотация

Представлена разработанная авторами объектно-ориентированная веб-система, предназначенная для формирования метаданных, описывающих результаты научной деятельности сотрудников учреждения (группы учреждений), и предоставления различных справочно-статистических данных о публикациях и докладах, сделанных ими на научных конференциях, симпозиумах, семинарах. Система ориентирована на работу с объектами таких связанных между собой классов, как «автор», «организация», «публикация», «доклад», «мероприятие». Профиль метаданных объектов каждого класса включает атрибуты, необходимые для получение развернутой информации как об отдельном объекте данного класса, так и о группе объектов, связанных заданными значениями атрибутов объектов других классов (например, перечень статей сотрудников заданного подразделения данной организации, опубликованных в заданном журнале за заданный промежуток времени). Отличительной особенностью системы является введенное понятие «эквивалентных» объектов. Эквивалентными считаются объекты, представленные в системе различными метаданными, но относящимися к одной физической сущности. Такими объектами являются «персоны», соответствующие одному автору с различными написаниями фамилии в библиографических описаниях публикаций; организации, имеющие различные варианты названий; статьи,

опубликованные без изменений на различных языках. Подробно охарактеризованы возможности системы, ее пользовательский интерфейс, приведены примеры выполнения конкретных запросов.

***Ключевые слова**: базы данных, учет результатов научной деятельности, веб-ориентированная система, сетевые технологии, анализ публикационной активности, программное обеспечение.*

ВВЕДЕНИЕ

Развиваемые в России подходы к оценке эффективности исследований, проводимых научными организациями, базируются в значительной степени на оценках публикационной активности их сотрудников. Появившиеся в последнее время новые критерии оценки эффективности научной деятельности опираются не только на количественную, но и на качественную оценку публикационной активности, в частности, на характеристики журнала, в котором опубликована та или иная статья.

Рост количества научных публикаций, возрастающие требования к многоаспектности отчетности, связанной с результатами интеллектуальной деятельности ученых, обусловливают необходимость создания информационно-справочных систем, позволяющих решать задачи в этой области.

В Российской академии наук задачи учета публикаций сотрудников традиционно решались библиотеками, обслуживающими тот или иной академический институт. Многие библиотеки в течение десятилетий вели картотеки трудов сотрудников, а с развитием вычислительной техники перешли к ведению соответствующих баз данных. Хотя базы данных публикаций сотрудников институтов играют важную роль в задачах наукометрии (что убедительно показано в [1]), единый подход к формированию таких баз данных отсутствует. Какие-то библиотеки или институты ведут базу данных публикаций сотрудников в EXCEL, какие-то – на основе собственного программного обеспечения, многие ограничиваются списками публикаций, представленными в виде текстовых файлов [2–6].

В современных условиях, когда для каждой организации большое значение приобретают вопросы учета публикаций (а шире – результатов интеллектуальной деятельности, включая доклады на научных конференциях, полученные патенты,

авторские свидетельства), представляется целесообразным разработать типовую систему, решающую общую задачу формирования и поддержки базы данных результатов интеллектуальной деятельности того или иного коллектива ученых. Подобная система разработана в МСЦ РАН с учетом опыта предыдущих разработок авторов [7, 8].

1. СТРУКТУРА СИСТЕМЫ

Система оперирует с 5-ю связанными классами объектов – «персона», «публикация», «организация», «доклад», «мероприятие». Класс «публикация» включает три связанных подкласса – публикация на аналитическом уровне (статья в журнале, сборнике); публикация на монографическом уровне (книга, брошюра, выпуск журнала и т. п.); публикация на сводном уровне (журнал, сборник). Для каждого класса разработан свой профиль метаданных (перечень обязательных и факультативных атрибутов) входящих в него объектов, и определены виды связей между объектами внутри класса и вне его.

Принципиальной особенностью данной системы, отличающей ее от других подобных, является введение связей между объектами типа «эквивалентные записи». Объекты, связанные таким образом, система воспринимает как одинаковые. Необходимость ввода эквивалентных персон обусловлена тем, что написание фамилии и имени одного автора в разных библиографических описаниях может отличаться. Например, в англоязычных публикациях персоны «Сотников А.Н.» можно встретить следующие написания данного автора: «Sotnikov A.N.», «Sotnikov A.», «A. Sotnikov», «A.N. Sotnikov», «Alexander Sotnikov». Не говоря о латинской транслитерации кириллических фамилий (одна фамилия может быть записана во многих вариантах), различные написания встречаются и в русскоязычных публикациях, например, при использовании буквы «е» и «ё», а также в таких именах, как «Наталья» и «Наталия», и т. п.

Эквивалентность публикаций возникает тогда, когда в базе данных отражены статья на языке оригинала и ее версии, полностью переведенные на другой язык. Связь эквивалентности для организаций устанавливается, когда организация, не изменяющаяся по сути, меняет свое название (например, ВЦ АН СССР и ВЦ

РАН). Благодаря наличию связей эквивалентности при обработке запроса, содержащего одно из значений того или иного атрибута, система выдаст результат, относящийся ко всем эквивалентным значениям данного атрибута.

В качестве примера приведем профили метаданных объектов классов «персона», «мероприятие» и подкласса «публикация на аналитическом уровне»; (о) после наименования атрибута означает, что атрибут обязательный, (ф) – факультативный.

Персона[1]:

- ✓ Фамилия и инициалы автора (о);
- ✓ Дополнительная информация (ф);
- ✓ Ссылки на организации (о);
- ✓ Ссылки на эквивалентные персоны (ф).

Мероприятие:

- ✓ Название мероприятия (о);
- ✓ Вид мероприятия (конференция, семинар и т. п.), выбирается из настраиваемого списка значений (о);
- ✓ Место проведения мероприятия (страна, город) (о);
- ✓ Сроки проведения мероприятия (представляются в нормализованном виде (дата начала – дата окончания): гггг.мм.дд – гггг.мм.дд (о);
- ✓ Адрес сайта мероприятия (ф);
- ✓ Адрес сайта с материалами мероприятия (ф);
- ✓ Дополнительная информация (ф).

Публикация на аналитическом уровне:

- ✓ Название публикации (о);
- ✓ Вид публикации (статья из журнала, сборника, тезисы докладов и т. п.), выбирается из настраиваемого списка значений (о);
- ✓ Год издания (о);
- ✓ Ссылка на объект монографического уровня (о)
- ✓ Страницы (о);

[1] Во избежание конфликта с законодательством об охране персональных данных при описании персоны используются минимальные сведения, идентифицирующие сотрудника данной организации.

- ✓ Адрес полного текста публикации (ф);
- ✓ Идентификаторы во внешних базах данных (в частности, DOI) (ф);
- ✓ Цитирование в WoS (ф);
- ✓ Цитирование в Scopus (ф);
- ✓ Цитирование в РИНЦ (ф);
- ✓ Ссылки на персон, являющихся авторами (о);
- ✓ Ссылки на эквивалентные публикации (ф);
- ✓ Дополнительная информация (ф).

Все операции, связанные с вводом и редактированием данных в системе, могут выполняться только авторизованными пользователями, поэтому наряду с вышеперечисленными в системе используются объекты вида «Оператор», содержащие информацию об авторизованных пользователях и их правах при работе с системой.

Права доступа делятся на две категории: права администратора системы и оператора. Администратор имеет право вводить и редактировать данные об операторах и редактировать все данные информационной базы системы. Оператор имеет возможность вводить и редактировать публикации сотрудников организации, которую он представляет.

Система состоит из двух модулей: административного (http://dirsmsc.ru/bd/adm.aspx) и пользовательского (http://dirsmsc.ru/bd/).

2. АДМИНИСТРАТИВНЫЙ МОДУЛЬ СИСТЕМЫ

В административном модуле осуществляются следующие процессы:

- ➢ ввод и редактирование данных об операторах, работающих с системой;
- ➢ ввод новых записей публикаций, докладов, мероприятий, персон, организаций;
- ➢ редактирование метаданных всех объектов;
- ➢ поиск и просмотр зарегистрированных в системе объектов;
- ➢ создание групп эквивалентных записей.

Рассмотрим процесс ввода в систему публикаций и докладов. Ввод данных новой публикации начинается с ввода ее авторов в том порядке, который представлен в публикации. Для каждого автора проверяется наличие его метаданных

в системе, в случае их отсутствия запускается процесс регистрации новой персоны. В систему вводятся данные автора (фамилия и инициалы, дополнительная информация), и формируется их связь с организацией, которая либо уже существует в системе, либо регистрируется как новая.

После окончания ввода авторов система открывает форму для ввода метаданных публикации: название, вид (статья, монография), источник, год издания, том, номер, страницы, адрес полного текста, идентификаторы во внешних базах данных (см. рис. 1). При вводе названия публикации система показывает уже зарегистрированные публикации (по совпадению авторов и первых слов заглавия). Для привязки к вводимой публикации источника (издания сводного уровня, где опубликована статья) его нужно найти по фрагментам названия, а в случае отсутствия – зарегистрировать в системе (ввести название и дополнительную информацию). После окончания ввода всех необходимых метаданных публикация будет зарегистрирована в системе.

Регистрация нового доклада так же, как и ввод новой публикации, начинается с ввода авторов. Затем система предоставляет форму для ввода метаданных доклада: название доклада, вид доклада (пленарный, секционный, стендовый, приглашенный), дополнительная информация. К докладу нужно привязать мероприятие, на котором был сделан доклад. По фрагментам названия мероприятия определяется его наличие в системе. В случае его отсутствия предоставляется форма для его ввода, включающая: вид мероприятия (конференция, семинар, симпозиум, совещание); место его проведения (город и страна); даты проведения мероприятия (начало и конец); ссылку на сайт мероприятия; ссылки на опубликованные материалы (см. рис. 2).

Следует отметить, что метаданные организаций представлены в системе в виде иерархической структуры: организация может включать подразделения, в которых есть отделы, которые, в свою очередь, могут включать лаборатории, и т. д. Администратор системы вводит название организации, затем названия ее подразделений, далее к каждому подразделению привязывает названия его отделов и т. д. На рис. 3 показан интерфейс для редактирования названий организаций на примере организации «Научно-исследовательский институт системных

исследований (НИИСИ РАН)». Здесь можно корректировать названия организации и подразделений, добавлять (или удалять) подразделения на любом уровне.

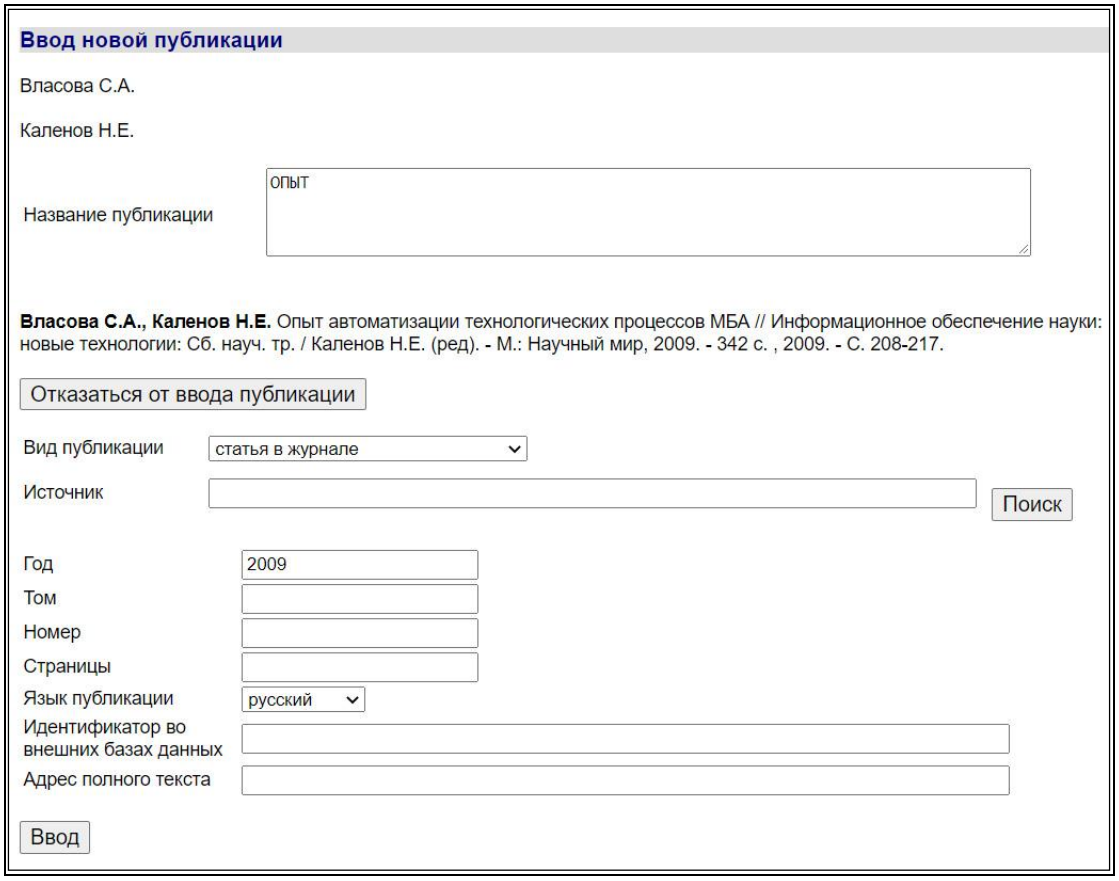


Рис. 1. Регистрация публикации

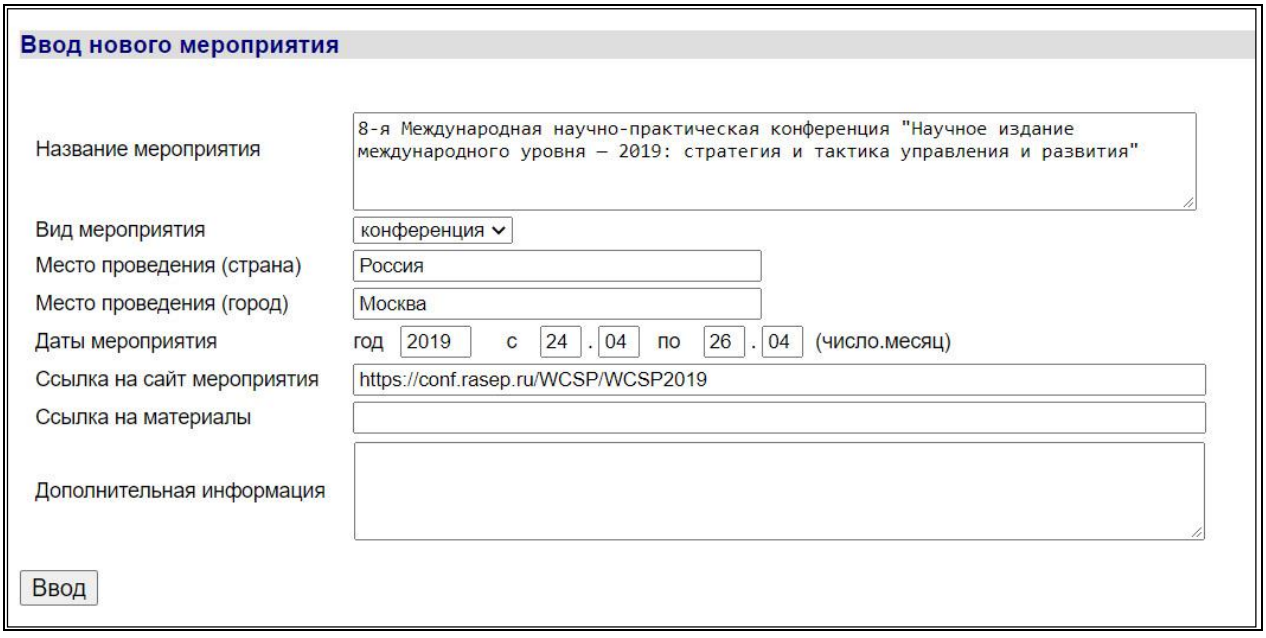


Рис. 2. Регистрация мероприятия

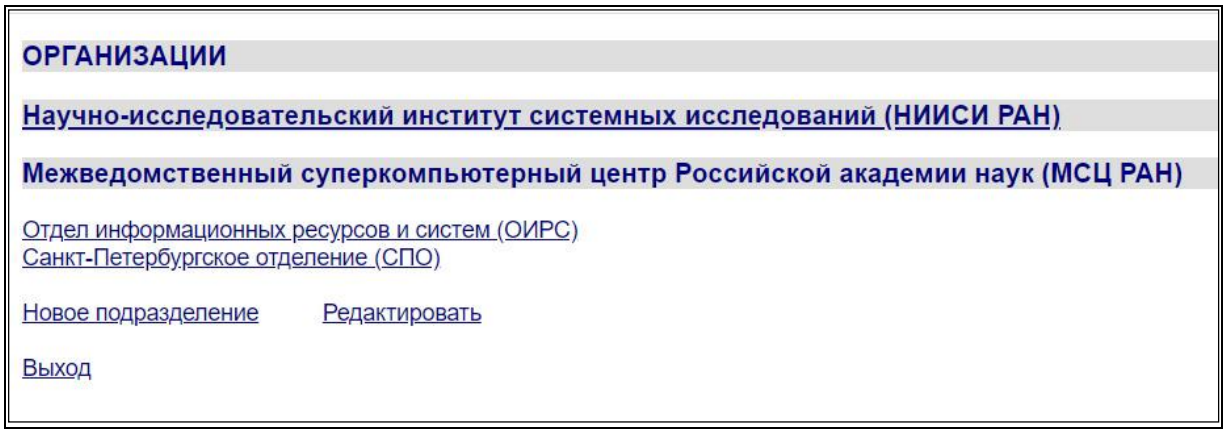


Рис. 3. Административный блок. Редактирование организации

3 ПОЛЬЗОВАТЕЛЬСКИЙ МОДУЛЬ СИСТЕМЫ

Пользовательский блок системы предоставляет возможность формировать многоаспектные запросы к системе, выдает на экран найденную информацию и обеспечивает навигацию по различным связанным объектам. При создании поискового интерфейса использовались подходы, реализованные авторами в других разработках, эффективность которых подтверждена на практике [10, 11]. Для формирования запроса в поисковую форму (см. рис. 4) вводятся термины в одну, две или три строки. Строки соединяются логическими операторами «И», «ИЛИ», «И НЕ». В каждую строку можно ввести несколько терминов, связав их логическими «И» или «ИЛИ». Возможно усечение термина справа, для этого используется символ «*».

Для каждой строки введенных терминов в поисковой форме выбирается наименование атрибута класса, в котором должен быть осуществлен поиск:

- ✓ Фамилия персоны
- ✓ Название публикации
- ✓ Название журнала / сборника
- ✓ Название доклада
- ✓ Мероприятие (название, страна, город)
- ✓ Организация (наименование, подразделение)
- ✓ Идентификатор во внешних базах данных

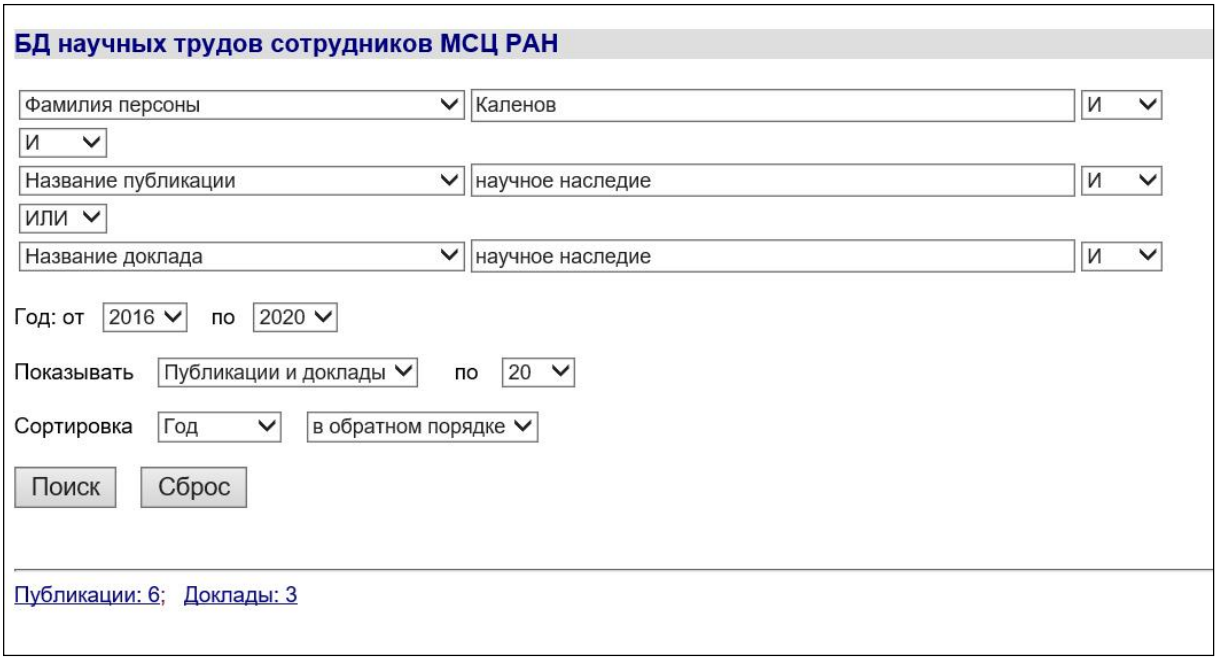


Рис. 4. Поисковая форма системы

По умолчанию, в поисковой форме в первой строке указано поле «Фамилия персоны», во второй – «Название публикации», в третьей – «Название доклада». Поисковый запрос можно ограничить годами публикаций (мероприятий), выбрав необходимые года из выпадающих списков «Год издания от … по …». Результаты поиска могут быть отсортированы по году или алфавиту описаний найденных объектов в прямом или обратном порядке. По умолчанию сортировка производится по году в обратном порядке (вначале выдаются публикации (мероприятия) текущего года). В поисковой форме системы имеется возможность настройки выдачи найденной информации путем выбора необходимой строки из выпадающего списка «Показывать»:

- ✓ Публикации и доклады
- ✓ Публикации
- ✓ Доклады
- ✓ Журналы / сборники
- ✓ Мероприятия
- ✓ Персоны
- ✓ Организации

По умолчанию, в поисковой форме выбрана опция «Публикации и доклады». Результат выполнения поисковых запросов выдается на экран порциями,

размер которых задается в выпадающем списке «Показывать … по …» (по умолчанию – 20 документов на странице). В том случае, если все поисковые поля оставить пустыми и нажать на кнопку «Поиск», система выдаст все зарегистрированные объекты в соответствии с выбранной опцией «Показывать».

Рассмотрим результаты обработки системой поисковых запросов при различных вариантах выбора показа найденных объектов.

Публикации и доклады

На рис. 4. приведен пример запроса на поиск публикаций и докладов автора «Каленов» за 2016–2020 гг., названия которых содержат термины «научное» и «наследие». По этому запросу система находит: «Публикации 6», «Доклады 3» (см. рис. 4). Переход по ссылке «Публикации …» открывает в новом окне браузера список публикаций; переход по ссылке «Доклады …» – список докладов (см. ниже).

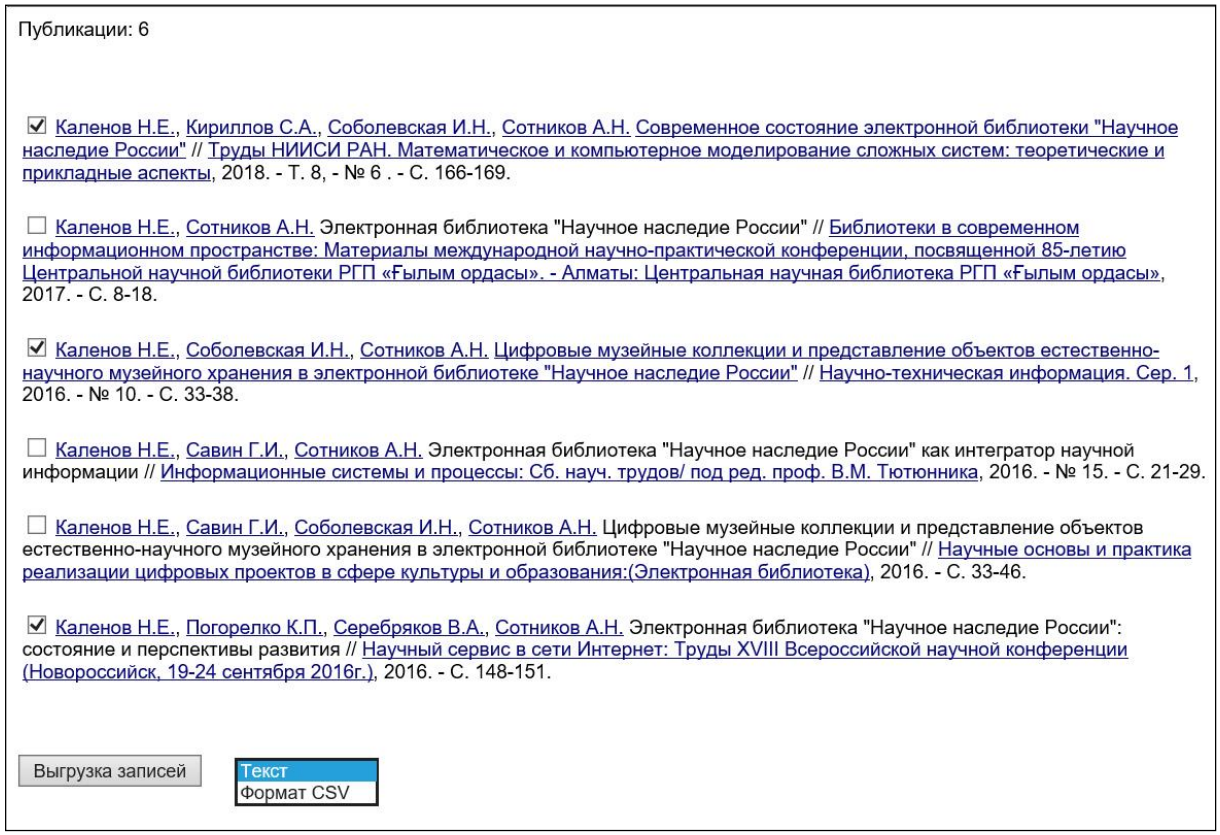


Рис. 5. Библиографические описания найденных публикаций

Публикации

Публикации выдаются в виде стандартных библиографических описаний, в которых авторы и названия журналов (сборников) являются активными ссылками (см. рис. 5). В том случае, если метаданные публикации содержат URL на полный текст публикации, название публикации также будет являться активной ссылкой, переход по которой обеспечит открытие статьи в новом окне браузера.

Переход по ссылке от фамилии автора обеспечит выдачу в новом окне браузера всех статей данного автора, зарегистрированных в системе. Кроме того, система покажет название организации, к которой относится автор, и относящуюся к нему дополнительную информацию. Ссылка от названия организации позволяет перейти на статьи всех персон, относящихся к данной организации.

В библиографическом описании публикации при переходе по ссылке от названия источника (журнала, сборника) пользователь получит в новом окне браузера описания всех статей, зарегистрированных в системе и опубликованных в данном источнике.

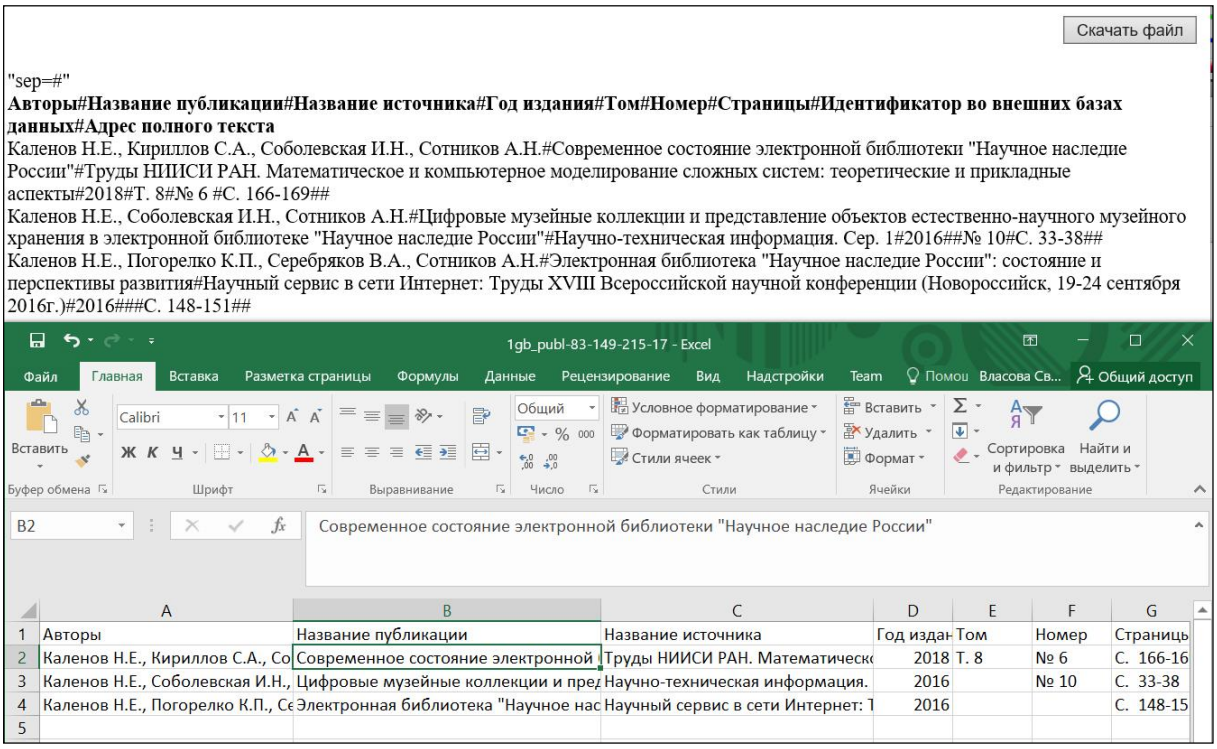


Рис. 6. Выгрузка публикаций в формате CSV

Система обеспечивает возможность выгрузки необходимых пользователю библиографических описаний стандартного вида в текстовый файл или в структурированный файл формата CSV. Записи файлов первого типа могут быть внесены в список пристатейной библиографии путем простого копирования, записи второго – в таблицу EXCEL для последующего использования в личной библиотеке пользователя (см. рис. 6).

Для выгрузки библиографических записей пользователь ставит «галочки» рядом с нужными ему публикациями, выбирает в нижней части страницы просмотра нужный формат и нажимает кнопку «Выгрузка записей». Отмечать записи можно на любой странице просмотра найденных публикаций, при переходе по страницам «галочки» сохраняются. Кнопку «Выгрузка записей» можно также нажать на любой из страниц. После ее нажатия в новом окне демонстрируется выгруженный файл выбранной структуры. Библиографические описания публикаций могут быть скопированы пользователем на свой компьютер простыми опциями copy / past. В случае выбора пользователем формата CSV на экране появляется кнопка «Скачать файл», после нажатия на которую записи выгружаются в файл с расширением CSV на ПК пользователя (см. рис. 6), при открытии которого в EXCEL формируется таблица со следующими столбцами:

- ✓ Авторы
- ✓ Название публикации
- ✓ Название источника
- ✓ Год издания
- ✓ Том
- ✓ Номер
- ✓ Страницы
- ✓ Идентификаторы во внешних базах данных
- ✓ Адрес полного текста
- ✓ Дополнительная информация

Доклады

Описания докладов содержат: авторов доклада, название доклада, описание мероприятия, на котором сделан доклад (см. рис. 7). Фамилии авторов представляют собой активные ссылки, переход по которым обеспечит выдачу в новом

окне браузера всех докладов данного автора, зарегистрированных в системе. Система покажет название организации, к которой относится автор, а также дополнительную информацию. Ссылка на названии организации позволяет перейти на описания докладов всех персон, относящихся к данной организации.

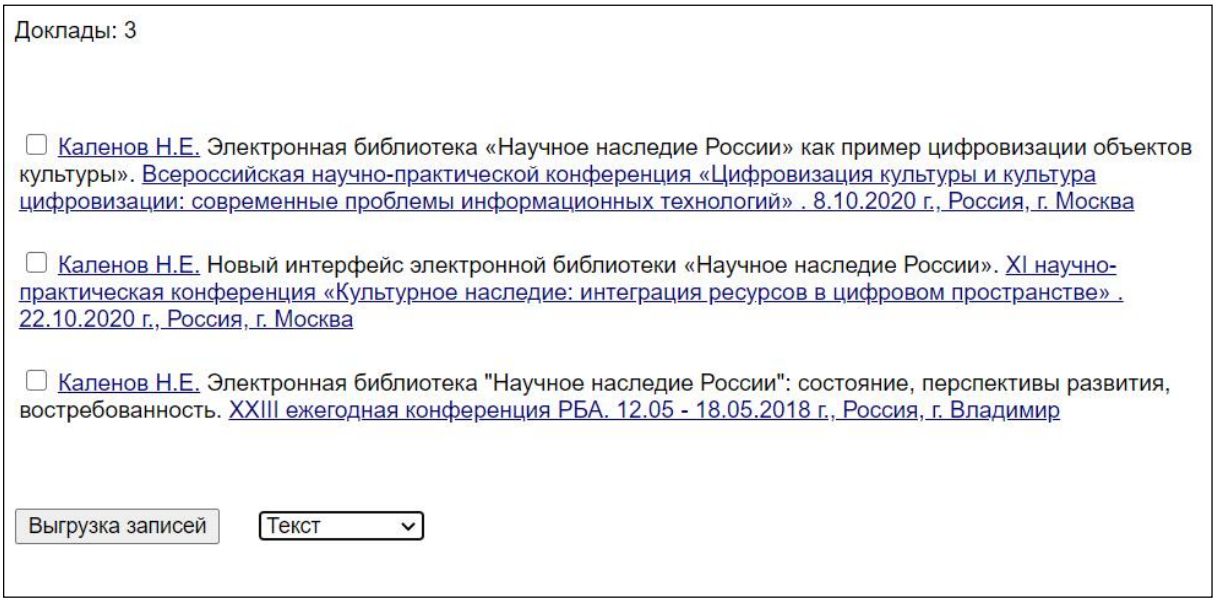


Рис. 7. Список найденных докладов

Если известен и введен в систему адрес сайта с дополнительной информацией о том или ином докладе, его название в списке (рис. 7) будет представлено в виде активной ссылки, по которой пользователь может перейти к тексту или презентации доклада. Ссылка от названия мероприятия обеспечивает переход на сайт мероприятия.

Система позволяет выгружать выбранные записи докладов в текстовом формате и формате CSV. Процедура отбора и выгрузки записей докладов полностью идентична описанной выше для выгрузки публикаций. При выгрузке в формате CSV формируется EXCEL-таблица со следующими столбцами:

- ✓ Авторы
- ✓ Название доклада
- ✓ Название мероприятия
- ✓ Год проведения
- ✓ Дата начала мероприятия
- ✓ Дата окончания мероприятия
- ✓ Страна

- ✓ Город
- ✓ Сайт мероприятия
- ✓ Сайт с материалами докладов

Журналы / сборники

В соответствии с обработанным поисковым запросом система выдает найденную информацию об изданиях на сводном уровне («источниках») – журналах, сборниках. Названия источников являются активными ссылками, переход по которым обеспечит выдачу в новом окне браузера всех публикаций, относящихся к выбранному источнику.

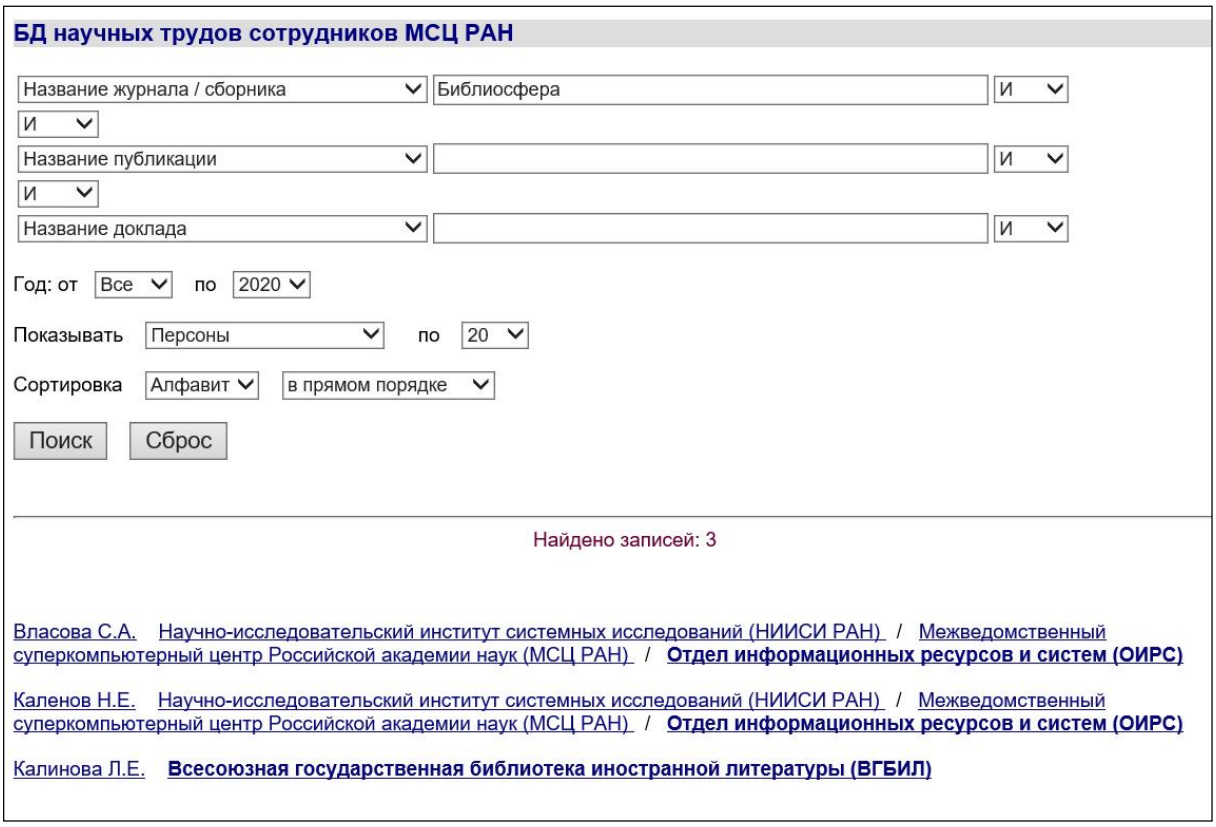


Рис. 8. Пример запроса на поиск персон

Мероприятия

Система выдает описания найденных мероприятий, в которых названия являются активными ссылками. При переходе по названию мероприятия в новом окне браузера получим список всех докладов, относящихся к выбранному мероприятию и зарегистрированных в системе.

Персоны

На рис. 8 приведен пример запроса на поиск персон, статьи которых были опубликованы в журнале «Библиосфера». Список, выводимый на экран, включает фамилии и инициалы найденных персон и названия связанных с ними организаций.

Фамилии персон представляют собой активные ссылки, при переходе по которым в новом окне браузера будет выдана все информация о выбранной персоне, имеющаяся в системе: название организации (активная ссылка), дополнительная информация, количество публикаций («Публикации …» – активная ссылка) и докладов («Доклады …» – активная ссылка). Переход по ссылкам «Публикации …» и «Доклады …» позволяет получить описания всех публикаций и докладов персоны. Переход по названию организации открывает новое окно с информацией о ней (см. ниже).

Организации

Метаданные организаций в системе представлены в виде иерархической структуры (организация может включать подразделения, например, отделы, которые, в свою очередь, могут включать лаборатории и т. д.). При выборе пользователем организаций в качестве результатов поиска система выдает список найденных объектов, названия которых являются активными ссылками. Если в результате поиска найдено подразделение организации, то система показывает не только его, но и все вышестоящие по иерархии подразделения данной организации. Например, для найденного подразделения «Отдел информационных ресурсов и систем (ОИРС)» (см. рис. 8) система покажет запись: *«Научно-исследовательский институт системных исследований (НИИСИ РАН) / Межведомственный суперкомпьютерный центр Российской академии наук (МСЦ РАН) / Отдел информационных ресурсов и систем (ОИРС)»*, в которой все названия, разделенные символом «/», являются активными ссылками. При переходе по ссылке выбранного названия система покажет все публикации и доклады, авторами которых являются сотрудники данного подразделения или организации в целом.

ЗАКЛЮЧЕНИЕ

Как показано выше, поисковый аппарат представленной системы обеспечивает пользователям удобную навигацию по связанным объектам: от найденной

статьи к источнику, в котором она опубликована, и далее ко всем статьям этого источника, зарегистрированным в системе; от найденного доклада к мероприятию, на котором он сделан, и далее ко всем докладам данного мероприятия; от автора к его публикациям и докладам; от организации к ее сотрудникам и ко всем их научным трудам.

В настоящее время система успешно работает в технологическом режиме в МСЦ РАН. Работу системы (ввод и редактирование записей) осуществляет один администратор. Пользователями системы являются научные сотрудники МСЦ РАН. В системе зарегистрированы 325 персон из 46 организаций; 743 статьи, опубликованные в 349 изданиях. Кроме того, в этом году начался ввод данных о докладах, сделанных на научных мероприятиях. Были зарегистрированы 25 мероприятий и 39 докладов, сделанных в 2018–2020 гг.

Благодарности

Работа выполнена в МСЦ РАН – филиале ФГУ ФНЦ НИИСИ РАН в рамках государственного задания № 0580-2021-0014.

СПИСОК ЛИТЕРАТУРЫ

1. *Мазов Н.А., Гуреев В.Н.* Библиографическая база данных трудов сотрудников организации: цели, функции, сфера использования в наукометрии // Вестник Дальневосточной государственной научной библиотеки. 2016. Вып. 2 (71). С. 84–87.

2. *Бескаравайная Е.В., Довбня Е.В., Захарова С.С.* Проблемно-ориентированные коллекции. Формирование и анализ на примере базы данных трудов сотрудников Института биофизики клетки // Библиография. 2008. № 4. С. 30–36.

3. *Левченко О.И., Соловьев А.В.* Формирование базы данных публикаций сотрудников Института физики твердого тела РАН // Информационное обеспечение науки: новые технологии: Сборник научных трудов. М.: БЕН РАН, 2015. С. 215–221.

4. Публикации сотрудников МИАН,

URL: http://www.mi-ras.ru/index.php?c=mianpubs&l=0&jrnfilters[]=jhep (дата обращения: 15.04.2020).

5. Публикации сотрудников Института Европы РАН,

URL: http://www.ieras-library.ru/a-z.htm (дата обращения: 15.04.2020).

6. *Королева И.Ю., Бахмад Э.А., Курочкина Е.В.* Картотека публикаций для ЭБС ВолгГТУ // Молодой ученый. 2012. № 6. С. 64–67.
URL: https://moluch.ru/archive/41/4894/ (дата обращения: 15.04.2020).

7. *Власова С.А., Каленов Н.Е.* Информатика в академической библиотеке // Системы и средства информатики. 2016. Т. 26. № 3. С. 162–178.

8. *Власова С.А.* Автоматизированная система поддержки корпоративной базы данных научных публикаций // Программные продукты, системы и алгоритмы. Электронный журнал. URL: http://www.swsys-web.ru, 2018. Вып. 2. С. 42–46.

9. *Власова С.А., Каленов Н.Е.* Новые поисковые возможности и востребованность каталога книг и продолжающихся изданий БЕН РАН // Информационное обеспечение науки: новые технологии: Сб. науч. тр. Екатеринбург, 2016. С. 171–178.

10. *Власова С.А., Каленов Н.Е.* Интернет-каталог Библиотеки по естественным наукам Российской академии наук как специальная информационно-поисковая система, ориентированная на квалифицированного пользователя // Системы и средства информатики. 2019. Т. 29. № 1. С. 86–95.

INFORMATION SYSTEM FOR REGISTERING THE RESULT OF SCIENTIFIC INSTITUTION EMPLOYEES' INTELLECTUAL ACTIVITY

S. A. Vlasova[1, [0000-0003-1533-5850]], N. E. Kalenov[2, [0000-0001-5269-0988]]

[1-2] Joint Supercomputer Center of the Russian Academy of Sciences – JSCC

[1]svlasova@jscc.ru, [2]nkalenov@jscc.ru

Abstract

The article describes a typical object-oriented WEB-system designed for storing and providing various reference and statistical data on the scientific works of employees of an institution (group of institutions), developed by specialists of the JSCC RAS. The system contains information about publications of employees and reports made by them at scientific conferences, symposiums, and seminars. The system is focused on working with objects belonged to classes connected between each other, such as

"author", "organization", "publication", "report", "event". The metadata profile of objects of each class includes attributes that are necessary to get detailed information about both an individual object of this class and a group of objects associated with the specified attribute values of objects of other classes. For example, you have to get a list of articles by employees of a given organization published articles in a given journal for a given period of time. A distinctive feature of the system is the introduced concept of "equivalent" objects. Such objects are "persons" corresponding to the same author with different spellings of the last name in the bibliographic descriptions of publications; organizations with different versions of names; articles which are published without changes in different languages. This article describes in detail the features of the system, its user interface, and provides examples of performing specific queries.

Keywords: *databases, research results accounting, WEB-based system, network technologies, publication activity analysis, software.*

REFERENCES

1. *Mazov N.A., Gureev V.N.* Bibliograficheskaya baza dannykh trudov sotrudnikov organizatsii: tseli, funktsii, sfera ispol'zovaniya v naukometrii // Vestnik Dal'nevostochnoy gosudarstvennoy nauchnoy biblioteki. 2016. Vyp. 2 (71). S. 84–87.

2. *Beskaravaynaya E.V., Dovbnya E.V., Zakharova S.S.* Problemno-orientirovannye kollektsii. Formirovanie i analiz na primere bazy dannykh trudov sotrudnikov Instituta biofiziki kletki // Bibliografiya. 2008. № 4. S. 30–36.

3. *Levchenko O.I., Solov'ev A.V.* Formirovanie bazy dannykh publikatsiy sotrudnikov Instituta fiziki tverdogo tela RAN // Informatsionnoe obespechenie nauki: novye tekhnologii: Sbornik nauchnykh trudov. M.: BEN RAN, 2015. S. 215–221.

4. Publikatsii sotrudnikov MIAN,
URL: http://www.mi-ras.ru/index.php?c=mianpubs&l=0&jrnfilters[]=jhep (accessed 17 November 2020).

5. Publikatsii sotrudnikov Instituta Evropy RAN,
URL: http://www.ieras-library.ru/a-z.htm (accessed 17 November 2020).

6. *Koroleva I.Yu., Bakhmad E.A., Kurochkina E.V.* Kartoteka publikatsiy dlya EBS VolgGTU // Molodoy uchenyy. 2012. № 6. S. 64–67.

7. *Vlasova S.A., Kalenov N.E.* Informatika v akademicheskoy biblioteke // Sistemy i sredstva informatiki, 2016. T. 26. № 3. S. 162–178.

8. *Vlasova S.A.* Avtomatizirovannaya sistema podderzhki korporativnoy bazy dannykh nauchnykh publikatsiy // Programmnye produkty, sistemy i algoritmy. Elektronnyy zhurnal. URL: http://www.swsys-web.ru, 2018. Vyp. 2. S. 42–46.

9. *Vlasova S.A., Kalenov N.E.* Novye poiskovye vozmozhnosti i vostrebovannost' kataloga knig i prodolzhayushchikhsya izdaniy BEN RAN // Informatsionnoe obespechenie nauki: novye tekhnologii: Sb. nauch. tr. Ekaterinburg, 2016. S. 171–178.

10. *Vlasova S.A., Kalenov N.E.* Internet-katalog Biblioteki po estestvennym naukam Rossiyskoy akademii nauk kak spetsial'naya informatsionno-poiskovaya sistema, orientirovannaya na kvalifitsirovannogo pol'zovatelya // Sistemy i sredstva informatiki, 2019. T. 29. № 1. S. 86–95.

СВЕДЕНИЯ ОБ АВТОРАХ

ВЛАСОВА Светлана Александровна – ведущий научный сотрудник Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», кандидат технических наук.

Svetlana Aleksandrovna VLASOVA – Leading Researcher of Joint Super Computer Center of the Russian Academy of Sciences – Branch of Federal State Institution «Scientific Research Institute for System Analysis of the Russian Academy of Sciences», Candidate of Technical Sciences.

email: svlasova@jscc.ru; ORCID: 0000-0003-1533-5850

КАЛЕНОВ Николай Евгеньевич – главный научный сотрудник Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», доктор технических наук, профессор.

Nikolay Evgenievich KALENOV – Chief Researcher of Joint Super Computer Center of the Russian Academy of Sciences – Branch of Federal State Institution «Scientific Research Institute for System Analysis of the Russian Academy of Sciences», Doctor of Technical Sciences, Professor.

email: nkalenov@jscc.ru; ORCID: 0000-0001-5269-0988

Материал поступил в редакцию 18 ноября 2020 года

УДК 004.4

АЛГОРИТМЫ ФОРМИРОВАНИЯ МЕТАДАННЫХ МАТЕМАТИЧЕСКИХ РЕТРО-КОЛЛЕКЦИЙ НА ОСНОВЕ АНАЛИЗА СТРУКТУРНЫХ ОСОБЕННОСТЕЙ ДОКУМЕНТОВ

П. О. Гафурова[1, [0000-0002-1544-155X]], **А. М. Елизаров**[2, [0000-0003-2546-6897]], **Е. К. Липачев**[3, [0000-0001-7789-2332]]

[1-3]Казанский (Приволжский) федеральный университет

[1]pogafurova@gmail.com, [2] amelizarov@gmail.com, [3] elipachev@gmail.com

Аннотация

Представлены решения основных задач, связанных с формированием цифровых математических коллекций из документов, изданных в доцифровой период, – такие коллекции обозначены в работе как ретро-коллекции. Приведены алгоритмы создания метаописания ретро-коллекций, основанные на анализе структуры математических документов и применении программных инструментов выделения метаданных. Дано описание ретро-коллекций, сформированных с помощью разработанных алгоритмов и включенных в состав фабрики метаданных цифровой математической библиотеки Lobachevskii-DML. Указаны схемы формирования метаданных и методы нормализации извлеченных метаданных в соответствии со схемами и требованиями интегрирующих математических библиотек.

Ключевые слова*: Lobachevskii-DML, фабрика метаданных, управление метаданными, цифровая ретро-коллекция.*

ВВЕДЕНИЕ

В Казанском университете, начиная с 2017 года, создается цифровая математическая библиотека Lobachevskii Digital Mathematical Library (Lobachevskii-DML, https://lobachevskii-dml.ru/) (см. также [1–3]). Одна из основных целей построения этой библиотеки состоит в разработке методов управления математическим контентом с помощью интеллектуальных программных инструментов. При проектировании архитектуры этой библиотеки учтены рекомендации, сформированные в рамках из-

вестных инициатив интеграции математического знания “World Digital Mathematics Library” (WDML, «Всемирная цифровая математическая библиотека», https://www.mathunion.org/ceic/library/world-digital-mathematics-library-wdml, 2012 год) и “The Global Digital Mathematics Library” (GDML, «Глобальная цифровая математическая библиотека», 2014 год) (см., например, [4–9]).

При создании методов управления метаданными математических цифровых коллекций нами применены форматы и схемы, реализованные в проекте “The European Digital Mathematics Library” – «Европейская Цифровая Математическая Библиотека» (EuDML, https://initiative.eudml.org/) (см., например, [10–14]). Использованы также подходы, реализованные в проекте MathNet.Ru (http://www.math-net.ru/), в рамках которого оцифрованы, снабжены метаданными и представлены в открытый доступ архивы многих российских математических научных журналов и других изданий (см., например, [15–20]).

В рамках проекта «Lobachevskii Digital Mathematical Library» разработана система взаимосвязанных программных инструментов, обеспечивающих создание, обработку, хранение, управление метаданными объектов цифровых библиотек и интеграцию создаваемых электронных коллекций в агрегирующие их цифровые научные библиотеки. Система таких инструментов составляет фабрику метаданных цифровой библиотеки (см. [21, 22]).

В настоящей работе рассмотрены методы, разработанные для формирования цифровых коллекций, содержащих математические документы, созданные в «доцифровой» период и существующие только в бумажном виде, – такие коллекции принято называть ретро-оцифрованными (retrodigitized) коллекциями (см., например, [23]). Далее мы будет использовать термин «ретро-коллекция», опуская слово «оцифрованный».

В разделе 1 описан процесс создания математических ретро-коллекций в цифровых библиотеках. Приведен ряд наиболее известных библиотек, осуществляющих оцифровку математических документов и формирования их метоописания.

В разделе 2 выделены основные методы управления контентом в цифровых математических библиотеках. Представлена система сервисов формирования и управ-

ления метаданными в цифровой математической библиотеке Lobachevskii-DML. Взаимосвязанная система таких сервисов, организованная в указанной библиотеке, названа фабрикой метаданных.

Третий раздел содержит описание методов анализа структуры документов, позволяющих найти информативные строки с последующим разбором и экстракцией метаданных.

В четвертом разделе отмечены структурные особенности математических ретро-коллекций, которые необходимо учитывать при поиске информации, необходимой для формирования набора метаданных. Описан процесс создания двух ретро-коллекций, включая автоматическое постатейное разделение номеров журналов, формирование обязательных наборов метаданных и последующее включение коллекций в состав цифровой математической библиотеки Lobachevskii-DML.

1. РЕТРО-КОЛЛЕКЦИИ В ЦИФРОВЫХ МАТЕМАТИЧЕСКИХ БИБЛИОТЕКАХ

Одна из целей Всемирной цифровой математической библиотеки (WDML) состоит в предоставлении в широкий доступ математических документов за весь период развития науки. В программных документах этого проекта отмечается, что математика является не только творческой (creative), но также и накопительной, совокупной, кумулятивной (cumulative) наукой (см. [4]). Кумулятивнось понимается в том смысле, что новые исследования всегда опираются на хорошо организованную (well-organized) и тщательно подобранную (well-curated) литературу. Отмечается также, что особенностью любого математического исследования являются исключительно логические рассуждения, без привязки к экспериментам. Математические документы рассматриваются как часть общей структуры математического знания. Можно даже утверждать, что математика является, пожалуй, единственной областью знаний, в которой цитирование почти никогда не является инструментом для противоречия (см., например, [24]). Поэтому для математиков важно, чтобы математическая литература была представлена и доступна в полном объеме.

Первоначальной задачей Цифровой математической библиотеки (DML) являлась ретрооцифровка, предполагающая создание цифровых копий документов, существующих только в бумажном виде. Одна из целей DML заключается в оцифровке всего существующего математического наследия. Процесс ретрооцифровки включает

также структурирование оцифрованной информации и формирование метаданных (см., например, [23]).

В работах [5, 6, 23] названы проекты, которые на протяжении многих лет реализуют идеи DML. Наиболее известными из них являются библиотека JSTOR (Journal STORage, https://www.jstor.org/), созданная в США в 1995 году [25, 26], французские библиотеки GALLICA (https://gallica.bnf.fr/) и NUMDAM (NUMérisation de Documents Anciens Mathématiques, http://www.numdam.org/), созданные, соответственно, в 1997 и 2000 годах (см., например, [24, 27, 28]), чешская библиотека DML-CZ (Czech Digital Mathematics Library, https://dml.cz/), развивающаяся с 2005 года (см., например, [29], [30]). В отличие от библиотек JSTOR и GALLICA, которые являются мультидисциплинарными, библиотеки NUMDAM и DML-CZ ориентированы на фундаментальную математику.

Наиболее значимым проектом по цифровизации математических документов на русском языке является «Общероссийский портал Math-Net.Ru», который развивается с 2006 года. В настоящее время на портале этого проекта (http://www.mathnet.ru/) в открытом доступе представлены цифровые коллекции российских математических журналов, начиная с момента их создания (см., например, [15–20]). Старейшим математическим журналом, выходящим на русском языке, является «Математический сборник», первый номер которого опубликован в 1866 году.

В настоящей работе описаны методы создания цифровых ретро-коллекций математических документов, хранящихся в Научной библиотеке им. Н.И. Лобачевского Казанского университета, и включения их в цифровую математическую библиотеку Lobachevskii-DML.

Отметим, что в Научной библиотеке им. Н.И. Лобачевского хранятся уникальные архивы математических документов 19–20 веков издания. Часть архивов находится в процессе оцифровки. Однако эти архивы не сформированы как цифровые коллекции с необходимыми для этого наборами метаданных и поисковыми сервисами.

2. УПРАВЛЕНИЕ КОНТЕНТОМ В ЦИФРОВОЙ МАТЕМАТИЧЕСКОЙ БИБЛИОТЕКЕ

Создание цифровой математической коллекции или библиотеки и последующее расширение её функциональных возможностей предполагают решение целого ряда трудоемких задач, связанных, в первую очередь, с управлением контентом. Именно поэтому программные инструменты управления научным контентом являются важнейшей составляющей любой цифровой библиотеки. Многие из этих инструментов используются фабрикой метаданных для создания, обработки, хранения и управления метаданными цифровых документов и позволяют интегрировать создаваемые электронные коллекции в агрегирующие цифровые научные библиотеки. Опишем подробнее имеющиеся решения.

Существующие цифровые библиотеки, а также агрегаторы научных знаний предлагают ряд программных инструментов для работы с контентом, прежде всего, сервисы поиска в электронных коллекциях. Например, средства семантического поиска документов представлены на сайте проекта EuDML (https://initiative.eudml.org/). Здесь же размещены демонстрационные версии инструментов, разработанных для обслуживания EuDML. Назначение и функциональные возможности этих программных инструментов описаны в [31].

Для оптимизации названных инструментов фабрики метаданных и последующей их модернизации было необходимо:

- определить особенности представления метаданных документов различных электронных коллекций, связанные как с применяемыми форматами, так и с изменениями состава и полноты набора метаданных в течение всего времени существования соответствующего научного издания;
- настроить программные инструменты управления научным контентом и адаптировать методы организации автоматизированной интеграции репозиториев математических документов с другими информационными системами;
- обеспечить нормализацию метаданных в соответствии с форматами агрегирующих библиотек.

В результате разработанными инструментами фабрики метаданных цифровой математической библиотеки Lobachevskii-DML стали (см. [21]):

- система сервисов автоматизированного формирования метаданных электронных математических коллекций;

- xml-язык представления метаданных, основанный на Journal Archiving and Interchange Tag Suite (NISO JATS) всех версий [32];
- созданные программные инструменты нормализации метаданных электронных коллекций научных документов в форматах, разработанных агрегаторами ресурсов по математике и Computer Science;
- алгоритм приведения метаданных к формату oai_dc и генерации структуры архивов для импорта в цифровое хранилище DSpace;
- методы интеграции существующих электронных математических коллекций Казанского университета в отечественные и зарубежные цифровые математические библиотеки [13, 14].

Первоначально документы проходят препроцессорную обработку, в результате которой выявляются файлы документов, обработка которых не поддерживается инструментами фабрики метаданных в автоматическом режиме. Для таких документов автоматически генерируется log-файл с отчетом. Эти файлы далее корректируются в ручном режиме.

Вместе с файлом документа в фабрику метаданных загружается справочная информация о документе, в частности, о его типе и кодировке. Основные документы, которые обрабатываются фабрикой метаданных, – это файлы статей в различных форматах. Поэтому одной из целей препроцессорной обработки является также определение типа документа: статья, монография или сборник статей. Дальнейшие действия выполняются для статей и монографий. Сборники статей разделяются программно (на основе структурных особенностей документа) на отдельные статьи, которые также отправляются на обработку в фабрику метаданных. Один из подходов к решению этой задачи описан в [33].

На этапе экстракции метаданных обрабатываются тексты документов с целью поиска обязательных метаданных (в терминологии [12]). Для этого используются шаблоны регулярных выражений и структурные особенности документов. Также на этом этапе производится исправление некоторых орфографических ошибок, возникающих при экстракции метаданных из текстов, полученных в результате распознавания

оцифрованных документов. Выполняются также исправление ошибочного выбора регистра и удаление лишних пробелов и знаков. Отметим, что этап экстракции является одним из базовых этапов функционирования фабрики метаданных.

Сервисы экстракции метаданных отвечают за извлечение метаданных из документов и внешних ресурсов. Извлечение основных метаданных на первом этапе экстракции существенно зависит от их наличия в документе в явном виде. Также для извлечения информации из текста применяются инструменты текстовой аналитики. В качестве внешних ресурсов могут использоваться коллекции цифровых документов, в которые входят рассматриваемая статья, а также интернет-ресурсы.

Широкое размещение в интернете метаданных различных документов привело к тому, что одним из их источников могут стать веб-страницы сайта-агрегатора метаданных или самой цифровой библиотеки. Таким образом, при формировании набора метаданных документов электронных коллекций, а также при получении дополнительных метаданных необходимо использовать метаданные, хранящиеся на внешних ресурсах. Эта задача сопряжена с задачами поиска информации в агрегирующих базах данных и цифровых библиотеках, некоторые из которых частично закрыты для доступа или прерывают соединение, позволяя скачивать только ограниченное количество метаданных. При поиске метаданных на страницах сайтов-агрегаторов нужно также учитывать, что выбор и порядок поиска в таких источниках должны быть определены заранее, так как некоторые источники хранят информацию только по конкретной тематике (например, библиографическая база данных DBLP – по компьютерной тематике) или же неполный список метаданных. Особенностями данного этапа является то, что к некоторым сайтам также ограничен режим доступа. Однако многие ресурсы предоставляют возможность легальной экстракции метаданных средствами API и сервера OAI-PMH. Основные шаги алгоритма экстракции метаданных из интернет-ресурса на примере одной из коллекций приведены в [13, 14].

На этапе верификации выполняется проверка полноты и соответствия состава выделенных метаданных установленным правилам, записанным в виде DTD-файлов или XML-схем. После прохождения этого этапа возможны три варианта дальнейших действий: повторные экстракция необходимых и дополнительных метаданных, а

также верификация; выдача отчета о том, что средства фабрики метаданных недостаточны для получения требуемого набора метаданных; переход к финальному этапу – нормализации метаданных.

Экстракция дополнительных метаданных направлена на извлечение мета-данных из источников, размещенных вне обрабатываемого документа. К таким источникам можно отнести коллекции, в которые входит обрабатываемый документ, а также интернет-ресурсы.

Ряд инструментов фабрики метаданных цифровой математической библиотеки разработан для выполнения процедур гармонизации и нормализации метаданных.

Гармонизация метаданных предполагает возможность одновременного использования нескольких различных стандартов метаданных в одной программной системе. С помощью методов нормализации метаданных выполняется отображение нескольких различных стандартов метаданных в единую схему или структуру для дальнейшего использования в единой программной системе (см., например, [13, 14, 34]).

Задачи, связанные с нормализацией метаданных в различные форматы, – одни из самых актуальных при работе фабрики метаданных. Примерами таких задач служат: нормализация в форматы для внутреннего хранения и загрузки в цифровую библиотеку; нормализация в форматы других цифровых библиотек и агрегаторов или представление в виде форматов библиографического цитирования.

3. МЕТОДЫ ЭКСТРАКЦИИ МЕТАДАННЫХ, ОСНОВАННЫЕ НА АНАЛИЗЕ СТРУКТУРЫ ДОКУМЕНТОВ

Научные документы, опубликованные в каком-либо периодическом издании, оформлены по правилам этого издания. В таких правилах определена четкая последовательность размещения структурных блоков документа: названия, предметных классификаторов, списка авторов, афилиации, аннотации, ключевых слов, списка литературы и приложений. Список шрифтов, используемых для оформления структурных блоков, также однозначно определен. Анализ структуры документов цифрового архива позволяет извлечь информацию об особенностях данного архива, разделить его на классы документов, схожих по структуре и оформлению, и разработать алгоритмы поиска строк для последующего извлечения из них метаданных. В таблице 1

приведен пример характерных признаков структурных блоков научной статьи, используемых для извлечения метаданных (подробнее см. [35, 36]). Для описания структуры научных документов разработаны специальные онтологии (см., например, [37, 38]). Для семантической структуризации цифрового контента в них используются онтологии CiTO, DoCo, SWAN, SKOS, CERIF и SPAR (см. [39, 40]).

Таблица 1.

Структурный блок	**Стилевые и структурные особенности блока**	**Концепт онтологии**
Title	Font: Times New Roman, 12 pt, bold, centered. Position: в начале документа	doco:title
Author's list	Font: Times New Roman, 12 pt, centered Position: после блока Title Regexp Pattern: И.О. Фамилия или И. Фамилия, перечисляются через запятую	doco:ListOfAuthors, feof:author
Affiliations	Font: Times New Roman, 12 pt, italic, centered Position: после Author's list	pro:relatesToOrganization
E-mail	Font: Times New Roman, 9 pt, bold, centered Position: после блока Affiliations Regexp Pattern: содержат символ @ и соответствует правилам URI	fabio:Email
Abstract	Font: Times New Roman, 9 pt Position: после блока E-mail Regexp Pattern: начинается со слов «Аннотация» или «Abstract».	doco: abstract
References	Posttion: в конце документа Regexp Pattern: начинается с заголовка «References», «Список литературы»	doco:bibliography, deo:BibliographicReference

Методы извлечения метаданных, основанные на анализе структуры документов и выявлении используемых стилевых правил, изложены в работах [41–44]. В статьях [35, 36] описан алгоритм автоматической обработки больших коллекций физико-математических документов, основанный на указанном подходе.

4. МЕТОД ФОРМИРОВАНИЯ РЕТРО-КОЛЛЕКЦИЙ В ЦИФРОВОЙ МАТЕМАТИЧЕСКОЙ БИБЛИОТЕКЕ LOBACHEVSKII-DML

Как первый пример формирования архивных коллекций опишем результаты создания цифровой коллекции «Трудов Математического центра им. Н.И. Лобачевского» (далее «Труды …»), полученные с помощью сервисов фабрики метаданных. В настоящее время оцифровано более 50 томов этого издания. «Труды …» издаются с 1998 года, и большинство томов содержит материалы математических конференций. Как следствие, большинство томов «Трудов …» состоит из несколько десятков статей с ограниченным (с современной точки зрения) составом метаданных. Кроме того, за период издания «Трудов …» было использовано несколько стилевых правил подготовки материалов, что отразилось на структуре документов и форматах файлов сформированных сборников. Необходимыми условиями создания цифровой коллекции из файлового массива «Трудов …» были разделение томов на отдельные статьи, выделение метаданных, описывающих каждую статью, генерация дополнительных метаданных (содержащих, в частности, библиографическое описание статьи, ссылку на файл статьи в цифровой коллекции, а также связи с профилями авторов статьи на академических порталах и в наукометрических базах данных (kpfu.ru, MathNet.ru, Scopus и др.). Разработанный алгоритм представлен в [33].

Также для загрузки метаданных в формате цифрового хранилища DSpace был создан сервис нормализации метаданных в соответствии со схемой oai_dc (см. Рис. 1). Сформированный сервис был апробирован на архиве «Трудов Математического центра им. Н.И. Лобачевского», а сформированная цифровая коллекция включена в состав цифровой библиотеки Lobachevskii DML (https://lobachevskii-dml.ru/journal/tmt).

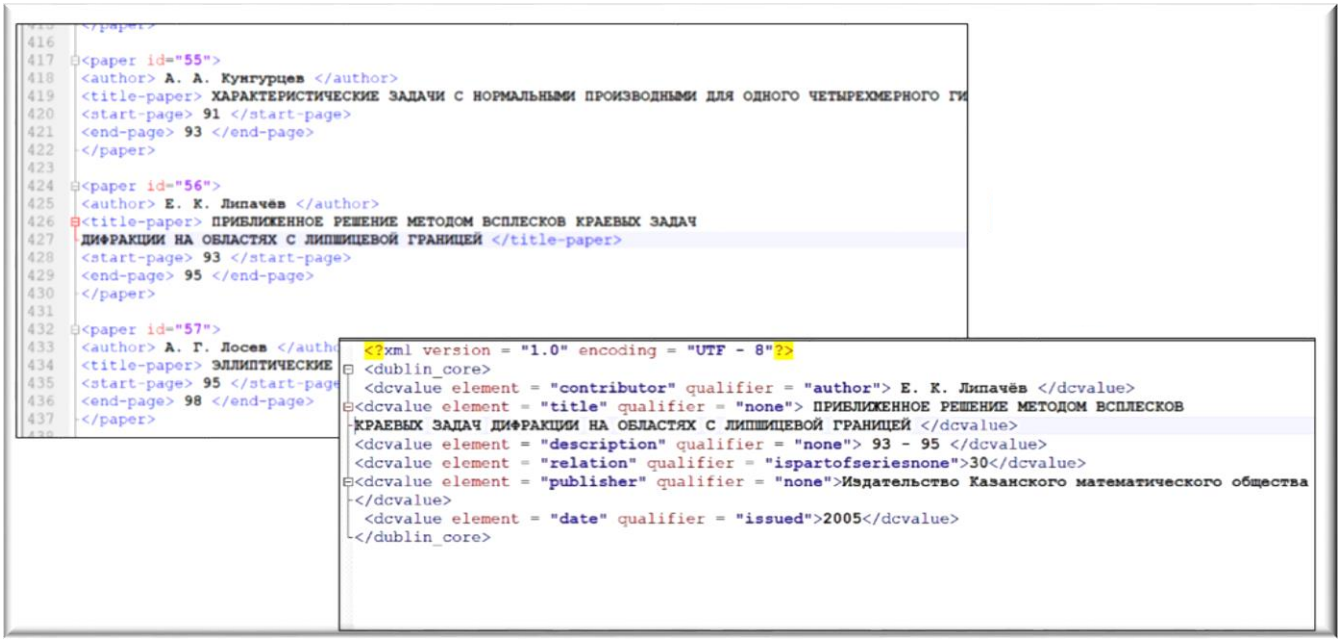


Рис. 1. Преобразование метаданных в спецификацию Dublin Core с учётом специфики цифрового хранилища DSpace.

Несомненный научный интерес представляет архив документов Физико-математического общества Казанского университета, который в настоящее время оцифрован лишь частично, а доступ к бумажным носителям ограничен. Его основой являются выпуски журнала «Известия физико-математического общества при Казанском университете» за 1891–1949 годы. В этом издании публиковались ведущие математики России, а позднее – Советского Союза. Среди авторов статей журнала – выдающиеся математики М.Г. Крейн, А.А. Марков, Н.Г. Чеботарёв и Н.Г. Четаев. Кроме того, опубликованы статьи и переводы работ Д. Гильберта, Ф. Клейна, С. Ли, А. Пуанкаре, Ш. Эрмита и других всем известных математиков.

Поскольку до момента формирования этой цифровой коллекции архив хранился только на бумажных носителях, необходимо было не только провести процедуры по экстракции метаданных документов коллекции, но и выполнить процесс оцифровки номеров журнала.

Выделим выполненные этапы формирования названной ретро–коллекции.

Этап 1. Создание метаописания архива статей журнала в форматах, допускающих машинную обработку. Предполагалось, что метаописание должно включать библиографическую запись всех статей указанного журнала. Поскольку журнал не оцифрован, этот этап автоматизировать не удалось. Дополнительное возникшее затруднение – необходимость работы с библиотечным бумажным фондом только с помощью системы выписок из каталога.

Этап 2. Представление цифровой коллекции в цифровой библиотеке Lobachevskii-DML в виде системы метаданных и ссылок на каталог Научной библиотеки Казанского университета.

Этап 3. Организация оцифровки архива указанного журнала.

Этап 4. Формирование цифровой коллекции, включающей полные тексты статей указанного журнала, снабженные наборами метаданных в форматах Lobachevskii-DML, MathNet.ru и формате обязательного набора метаданных «Европейской Цифровой Математической Библиотеки» (EuDML, https://initiative.eudml.org/).

Этап 5. Включение сформированной цифровой коллекции в Lobachevskii-DML с набором метаданных и полными текстами статей.

Отметим основные особенности рассматриваемого цифрового архива.

В зависимости от года издания сборники материалов архива имеют различное стилевое оформление статей. При этом в них практически отсутствует информация, необходимая для формирования фундаментального набора метаданных по схеме EuDML [10]. Трудности выделения метаданных из статей иллюстрируются рисунками 2 и 3. На рис. 2 показано, как оформлены статьи архива. На первой странице приводится название статьи, а на последней – автор. Это знаменитая статья А.А. Маркова «Распространение закона больших чисел на величины, зависящие друг от друга», опубликованная в номере 4 за 1906 год «Известий физико-математического общества при Казанском университете». В этой статье исследованы последовательности случайных событий, которые в настоящее время принято называть марковскими цепями.

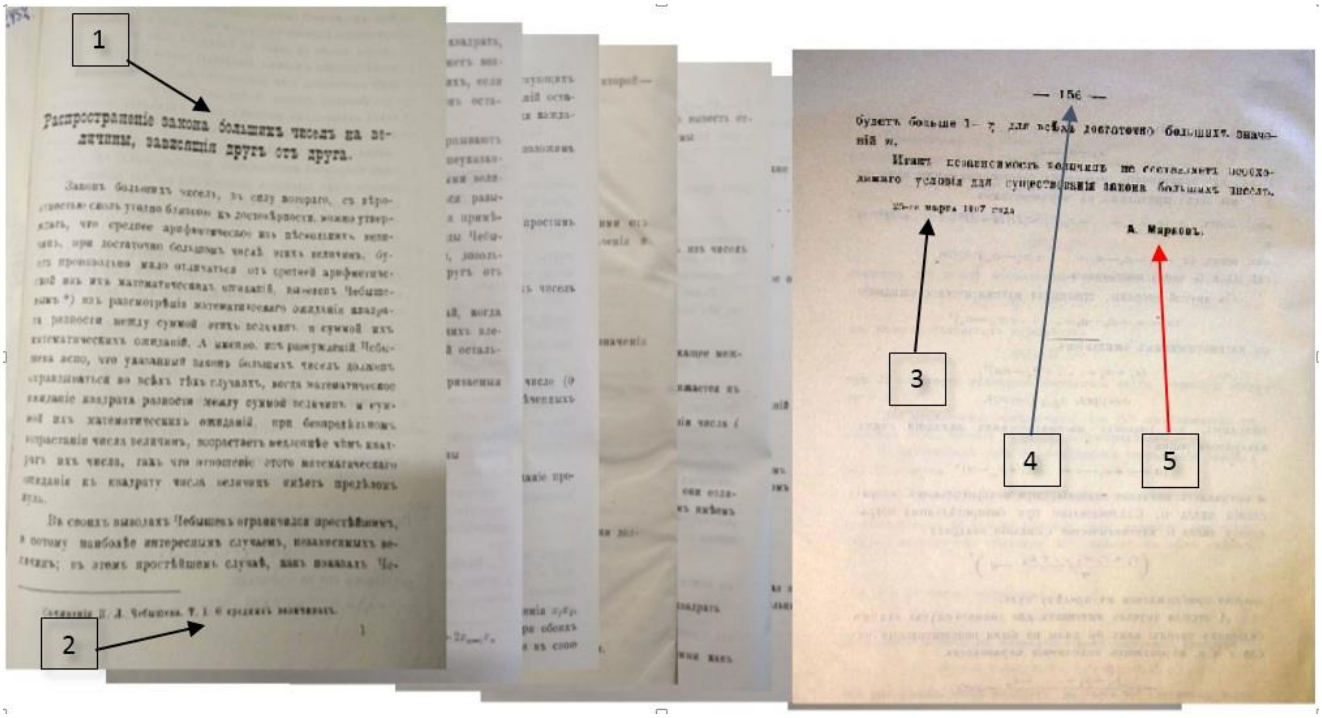


Рис. 2. Отсутствие в статьях информации, необходимой для формирования набора метаданных, в частности, сведений об авторах, ключевых слов. Структурные и стилевые особенности статей позволяют найти строки, из которых можно извлечь метаданные: 1 – название статьи, 2 – ссылка на научную статью или книгу, 3 – дата поступления статьи в редакцию журнала, 4 – номер завершающей страницы статьи, 5 – фамилия автора.

Программная обработка статей ретро–коллекции и формирование метаданных проводились в соответствии со следующим алгоритмом.

Алгоритм 1: Экстракция и нормализация метаданных статей второй серии «Известий….»

1: **читать** файл номера журнала в формате pdf
2: **загрузить** шаблон, определяющий структурные особенности номера
3: **вычислить** диапазоны страниц статей номера
4: **разделить** файл номера на файлы статей
5: **выделить** первую страницу статьи
6: **осуществить поиск строки** с названием статьи
7: **определить** основной язык статьи
8: **выделить** название статьи по шрифтовому шаблону
9: **преобразовать** название статьи в метаданные
10: **выделить** последнюю страницу статьи

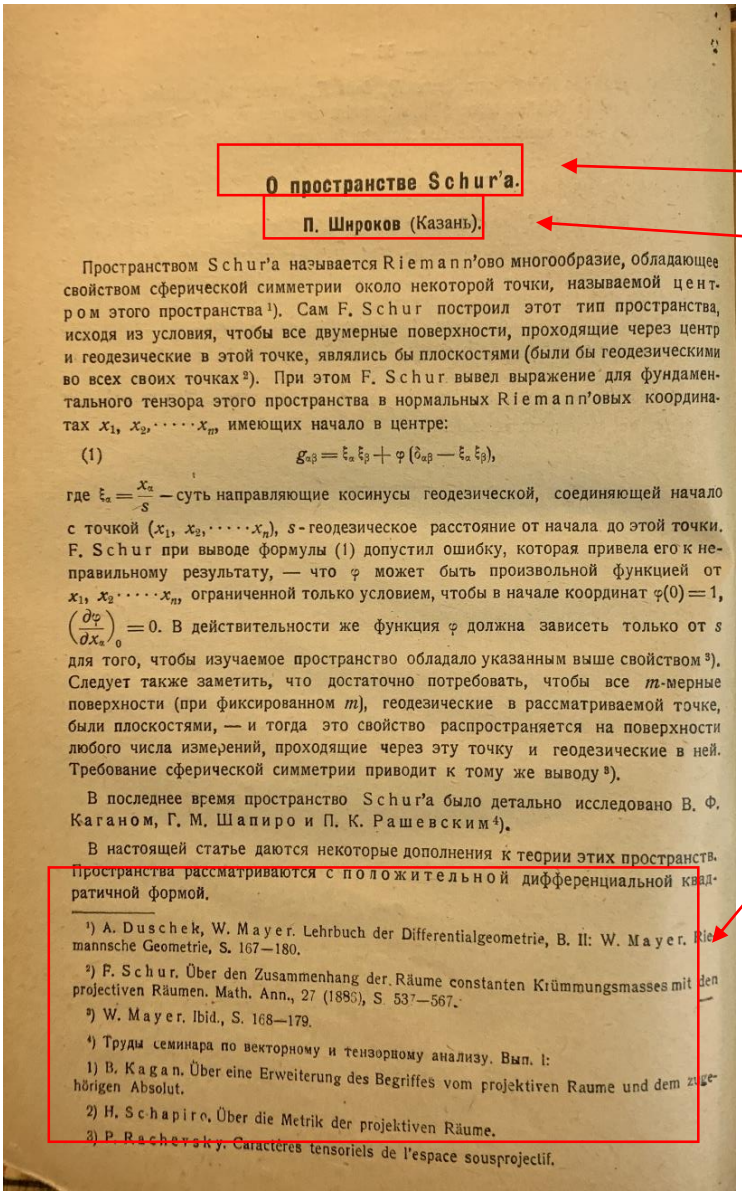
11: **осуществить поиск строки,** содержащей список авторов
12: **выделить** авторов статьи по шаблонам регулярных выражений
13: **осуществить поиск и извлечение** блока аннотации
14: **осуществить поиск и извлечение** списка литературы
15: **уточнить информацию об авторе** из открытых интернет источников
16: **сформировать** набор метаданных в соответствии схеме нормализации

На Рис. 3 приведен фрагмент набора метаданных, сформированный для статьи, представленной на Рис. 2. В журнале для этой статьи был приведен также перевод названия статьи и фамилии автора на французский язык – эта информация включена в метаданные. В метаописание включено также название, переведённое на современный русский язык, а также произведена процедура уточнения автора статьи с добавлением ссылки на статью об авторе, найденную в Сети.

```
<article id>2-15-4-1</article id>
<title-group>
    <article-title xml:lang="ru">Распространение закона больших чисел на
    величины, зависящие друг от друга.</article-title>
    <alt-title xml:lang="ru-o">Распространеніе закона большихъ чиселъ на
    величины, зависящія другъ отъ друга.</alt title>
    <alt-title xml:lang="fr">Extension de la loi de grands nombres aux
    événements dependants les uns des autres.</alt-title>
</title group>
<contrib-group>
    <contrib contrib-type="author">
        <name alternatives>
            <name>
                <surname xml:lang="ru">Марковъ</surname>
                <given names xml:lang="ru">А. А.</given names>
                <string-name xml:lang="ru-o">А. А. Марковъ </string-name>
                <string-name xml:lang="fr">A. Markof </string-name>
            </name>
            <uri>https://ru.wikipedia.org/wiki/
            Марков,_Андрей_Андреевич_(старший)</uri>
        </name alternatives>
    </contrib>
</contrib-group>
```

Рис. 3. Метаописание статьи ретро-коллекции, приведенной на Рис. 2.

Для статей из 3-й серии (и томов 23–25 из 2-й серии) характерен другой формат: языки публикаций – русский, немецкий, английский. Отметим, что фамилии зарубежных ученых в ссылках и названиях теорем в тексте статей не переводятся на русский язык (см. Рис. 4).



О пространстве Schur'a.

П. Широков (Казань).

Пространством Schur'a называется Riemann'ово многообразие, обладающее свойством сферической симметрии около некоторой точки, называемой центром этого пространства [1]). Сам F. Schur построил этот тип пространства, исходя из условия, чтобы все двумерные поверхности, проходящие через центр и геодезические в этой точке, являлись бы плоскостями (были бы геодезическими во всех своих точках [2]). При этом F. Schur вывел выражение для фундаментального тензора этого пространства в нормальных Riemann'овых координатах $x_1, x_2, \cdots x_n$, имеющих начало в центре:

$$(1) \qquad g_{\alpha\beta} = \xi_\alpha \xi_\beta + \varphi(\delta_{\alpha\beta} - \xi_\alpha \xi_\beta),$$

где $\xi_\alpha = \frac{x_\alpha}{s}$ — суть направляющие косинусы геодезической, соединяющей начало с точкой $(x_1, x_2, \cdots x_n)$, s - геодезическое расстояние от начала до этой точки. F. Schur при выводе формулы (1) допустил ошибку, которая привела его к неправильному результату, — что φ может быть произвольной функцией от $x_1, x_2 \cdots x_n$, ограниченной только условием, чтобы в начале координат $\varphi(0) = 1$, $\left(\frac{\partial \varphi}{\partial x_\alpha}\right)_0 = 0$. В действительности же функция φ должна зависеть только от s для того, чтобы изучаемое пространство обладало указанным выше свойством [3]). Следует также заметить, что достаточно потребовать, чтобы все m-мерные поверхности (при фиксированном m), геодезические в рассматриваемой точке, были плоскостями, — и тогда это свойство распространяется на поверхности любого числа измерений, проходящие через эту точку и геодезические в ней. Требование сферической симметрии приводит к тому же выводу [3]).

В последнее время пространство Schur'a было детально исследовано В. Ф. Каганом, Г. М. Шапиро и П. К. Рашевским [4]).

В настоящей статье даются некоторые дополнения к теории этих пространств. Пространства рассматриваются с положительной дифференциальной квадратичной формой.

[1]) A. Duschek, W. Mayer. Lehrbuch der Differentialgeometrie, B. II: W. Mayer. Riemannsche Geometrie, S. 167—180.

[2]) F. Schur. Über den Zusammenhang der Räume constanten Krümmungsmasses mit den projectiven Räumen. Math. Ann., 27 (1886), S. 537—567.

[3]) W. Mayer. Ibid., S. 168—179.

[4]) Труды семинара по векторному и тензорному анализу. Вып. I:

1) B. Kagan. Über eine Erweiterung des Begriffes vom projektiven Raume und dem zugehörigen Absolut.

2) H. Schapiro. Über die Metrik der projektiven Räume.

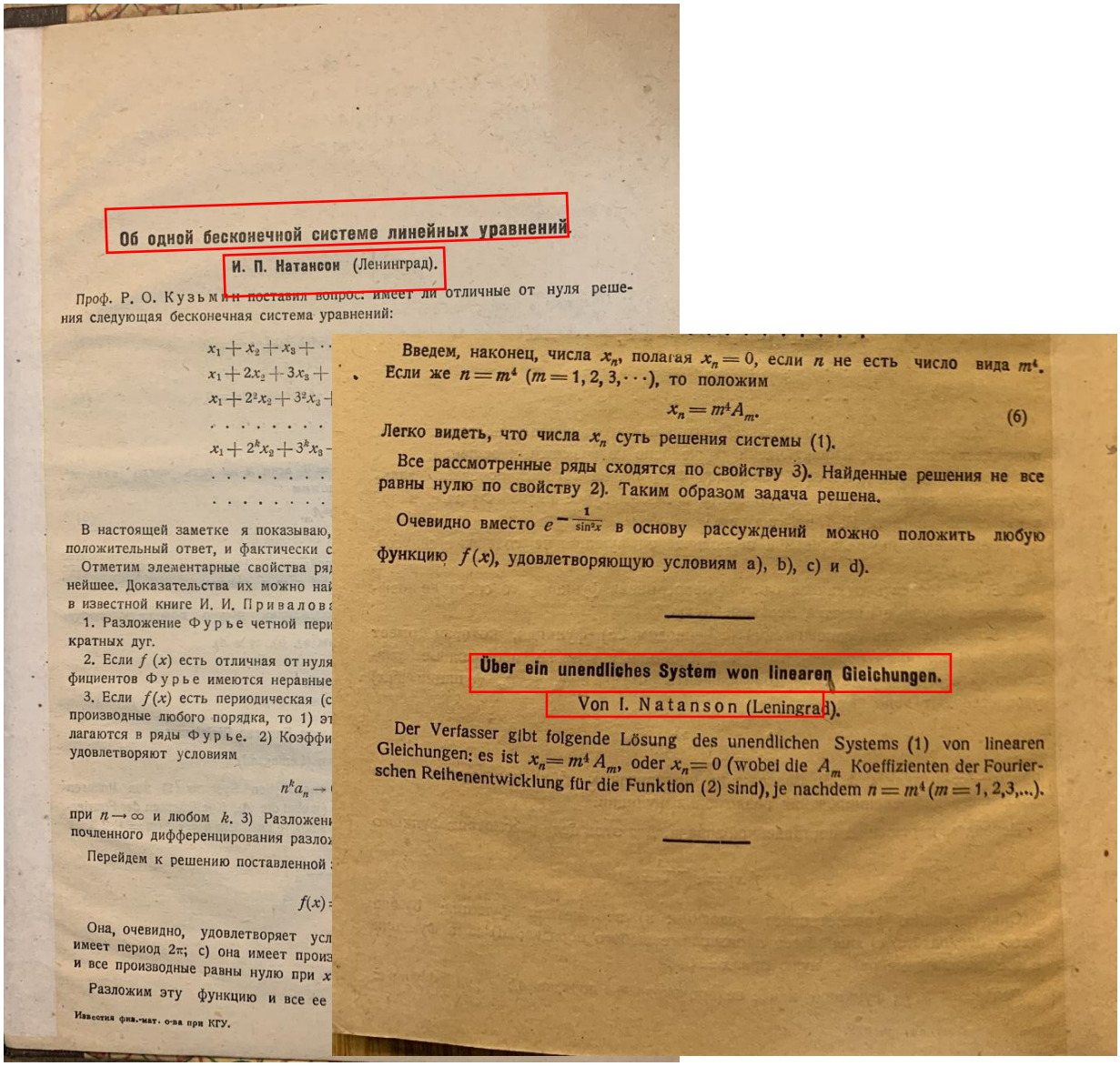
3) P. Rachevsky. Caractères tensoriels de l'espace sousprojectif.

Название

Фамилия автора и город проживания

Ссылки на научные источники

Рис. 4. Структурные и стилевые особенности статей из третьей серии. Название и список авторов указаны на первой странице статьи. После фамилии автора указан город его проживания. Список литературы не выделен в отдельный структурный блок, научные источники, используемые в статье, оформлены в виде сносок. В качестве примера приведена статья П.А. Широкова, опубликованная в третьей серии «Известий физико-математического общества при Казанском университете им. В.И. Ульянова-Ленина» (Широков П. О пространстве Schur'a // Изв. физ-мат. об-ва. 1934–1935. 7. С. 64–76).



Об одной бесконечной системе линейных уравнений.

И. П. Натансон (Ленинград).

Проф. Р. О. Кузьмин поставил вопрос: имеет ли отличные от нуля решения следующая бесконечная система уравнений:

$$x_1 + x_2 + x_3 + \cdots$$
$$x_1 + 2x_2 + 3x_3 + \cdots$$
$$x_1 + 2^2x_2 + 3^2x_3 + \cdots$$
$$\cdots\cdots\cdots$$
$$x_1 + 2^kx_2 + 3^kx_3 + \cdots$$
$$\cdots\cdots\cdots$$
$$\cdots\cdots\cdots$$

В настоящей заметке я показываю, положительный ответ, и фактически с

Отметим элементарные свойства ря нейшее. Доказательства их можно на в известной книге И. И. Привалов

1. Разложение Фурье четной пери кратных дуг.

2. Если $f(x)$ есть отличная от нуля фициентов Фурье имеются неравные

3. Если $f(x)$ есть периодическая (с производные любого порядка, то 1) эт лагаются в ряды Фурье. 2) Коэффи удовлетворяют условиям

$$n^k a_n \to$$

при $n \to \infty$ и любом k. 3) Разложени почленного дифференцирования разло

Перейдем к решению поставленной

$$f(x) =$$

Она, очевидно, удовлетворяет усл имеет период 2π; c) она имеет произ и все производные равны нулю при x

Разложим эту функцию и все ее

Известия физ.-мат. о-ва при КГУ.

Введем, наконец, числа x_n, полагая $x_n = 0$, если n не есть число вида m^4. Если же $n = m^4$ $(m = 1, 2, 3, \cdots)$, то положим

$$x_n = m^4 A_m. \qquad (6)$$

Легко видеть, что числа x_n суть решения системы (1).

Все рассмотренные ряды сходятся по свойству 3). Найденные решения не все равны нулю по свойству 2). Таким образом задача решена.

Очевидно вместо $e^{-\frac{1}{\sin^2 x}}$ в основу рассуждений можно положить любую функцию $f(x)$, удовлетворяющую условиям a), b), c) и d).

Über ein unendliches System won linearen Gleichungen.

Von I. Natanson (Leningrad).

Der Verfasser gibt folgende Lösung des unendlichen Systems (1) von linearen Gleichungen: es ist $x_n = m^4 A_m$, oder $x_n = 0$ (wobei die A_m Koeffizienten der Fourierschen Reihenentwicklung für die Funktion (2) sind), je nachdem $n = m^4 (m = 1, 2, 3, \ldots)$.

Рис. 5. На последней странице каждой статьи третьей серии приведен перевод на один из европейских языков (как правило, на немецкий) названия статьи, фамилий авторов и аннотации. В качестве примера приведена статья И.П. Натансона, опубликованная в третьей серии «Известий …» (Натансон И.П. Об одной бесконечной системе линейных уравнений // Изв. физ.-матем. о-ва при Казанском ун-те. 1934–1935. 7. С. 97–98).

Алгоритм 2: Экстракция и нормализация метаданных статей третьей серии «Известий ….»

1: **читать** файл номера журнала в формате pdf
2: **загрузить** шаблон, определяющий структурные особенности номера
3: **вычислить** диапазоны страниц статей номера
4: **разделить** файл номера на файлы статей
5: **выделить** первую страницу статьи
6: **осуществить поиск строки** с названием статьи
7: **определить** основной язык статьи
8: **выделить** название статьи по шрифтовому шаблону
9: **преобразовать** название статьи в метаданные
10: **осуществить поиск строки,** содержащей список авторов
11: **выделить** авторов статьи по шаблонам регулярных выражений
12: **осуществить поиск и извлечение** списка литературы
13: **выделить** вторую страницу статьи
14: **определить и вычислить** номер первой страницы статьи
15: **выделить** последнюю страницу статьи
16: **осуществить поиск аннотации** (если основной язык русский)
17: **определить** язык аннотации
18: **выделить** переводное название
19: **выделить** перевод имени автора и аффилиацию
20: **выделить** номер последней страницы
21: **уточнить информацию об авторе** из открытых интернет источников
22: **сформировать** набор метаданных в соответствии схеме нормализации

Ряд номеров второй серии «Известий физико-математического общества при Казанском университете» помимо научных статей содержит переводы на русский язык научных мемуаров известных математиков, конспекты лекций, формулировки нерешенных математических задач, а также письма ученых и новости мирового математического сообщества. Разнообразие типов материалов, как следствие, потребовало проведения кластеризации документов цифрового архива с целью выделения сходства по структуре и стилю.

Автор

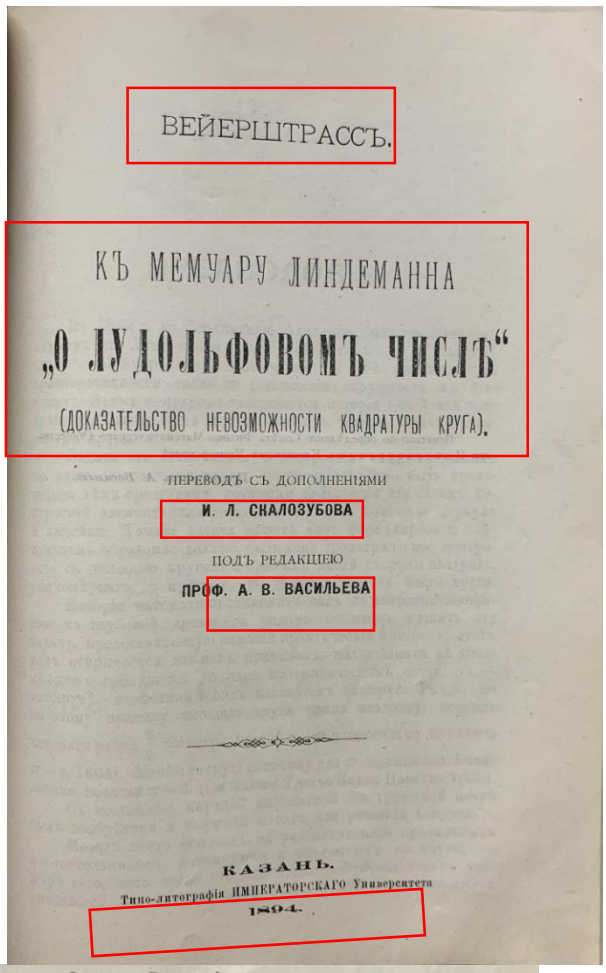
Название статьи

Переводчик

Редактор

Издательство

Год издания



ВЕЙЕРШТРАССЪ.

КЪ МЕМУАРУ ЛИНДЕМАННА

„О ЛУДОЛЬФОВОМЪ ЧИСЛѢ“

(ДОКАЗАТЕЛЬСТВО НЕВОЗМОЖНОСТИ КВАДРАТУРЫ КРУГА).

ПЕРЕВОДЪ СЪ ДОПОЛНЕНІЯМИ

И. Л. СКАЛОЗУБОВА

ПОДЪ РЕДАКЦІЕЮ

ПРОФ. А. В. ВАСИЛЬЕВА

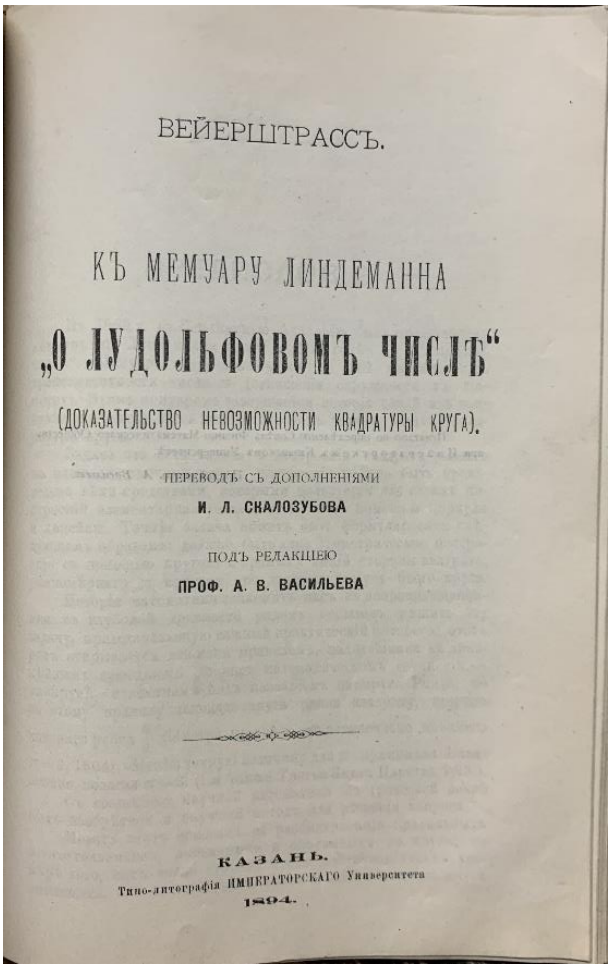
КАЗАНЬ.

Типо-литографія ИМПЕРАТОРСКАГО Университета

1894.

Рис. 6. Извлечение метаданных с использованием структурных особенностей. Отметим, что в этом случае необходимо провести процедуру уточнения метаданных, поскольку автор статьи указан не полностью, а переводчик и редактор приведены в родительном падеже.

На рис. 7 представлены некоторые типы материалов: документ (Устав физико-математического общества), статьи в переводе, письма.



ВЕЙЕРШТРАССЪ.

КЪ МЕМУАРУ ЛИНДЕМАННА

„О ЛУДОЛЬФОВОМЪ ЧИСЛѢ"

(ДОКАЗАТЕЛЬСТВО НЕВОЗМОЖНОСТИ КВАДРАТУРЫ КРУГА).

ПЕРЕВОДЪ СЪ ДОПОЛНЕНІЯМИ

И. Л. СКАЛОЗУБОВА

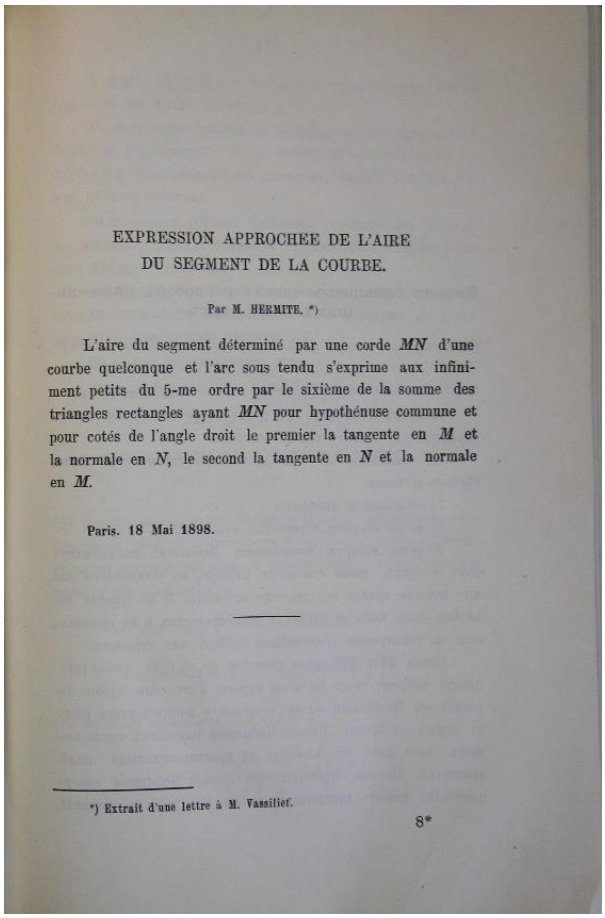
ПОДЪ РЕДАКЦІЕЮ

ПРОФ. А. В. ВАСИЛЬЕВА

КАЗАНЬ.

Тип.-литография ИМПЕРАТОРСКАГО Университета

1894.



EXPRESSION APPROCHEE DE L'AIRE
DU SEGMENT DE LA COURBE.

Par M. HERMITE. *)

L'aire du segment déterminé par une corde *MN* d'une courbe quelconque et l'arc sous tendu s'exprime aux infiniment petits du 5-me ordre par le sixième de la somme des triangles rectangles ayant *MN* pour hypothénuse commune et pour cotés de l'angle droit le premier la tangente en *M* et la normale en *N*, le second la tangente en *N* et la normale en *M*.

Paris. 18 Mai 1898.

*) Extrait d'une lettre à M. Vassilief.

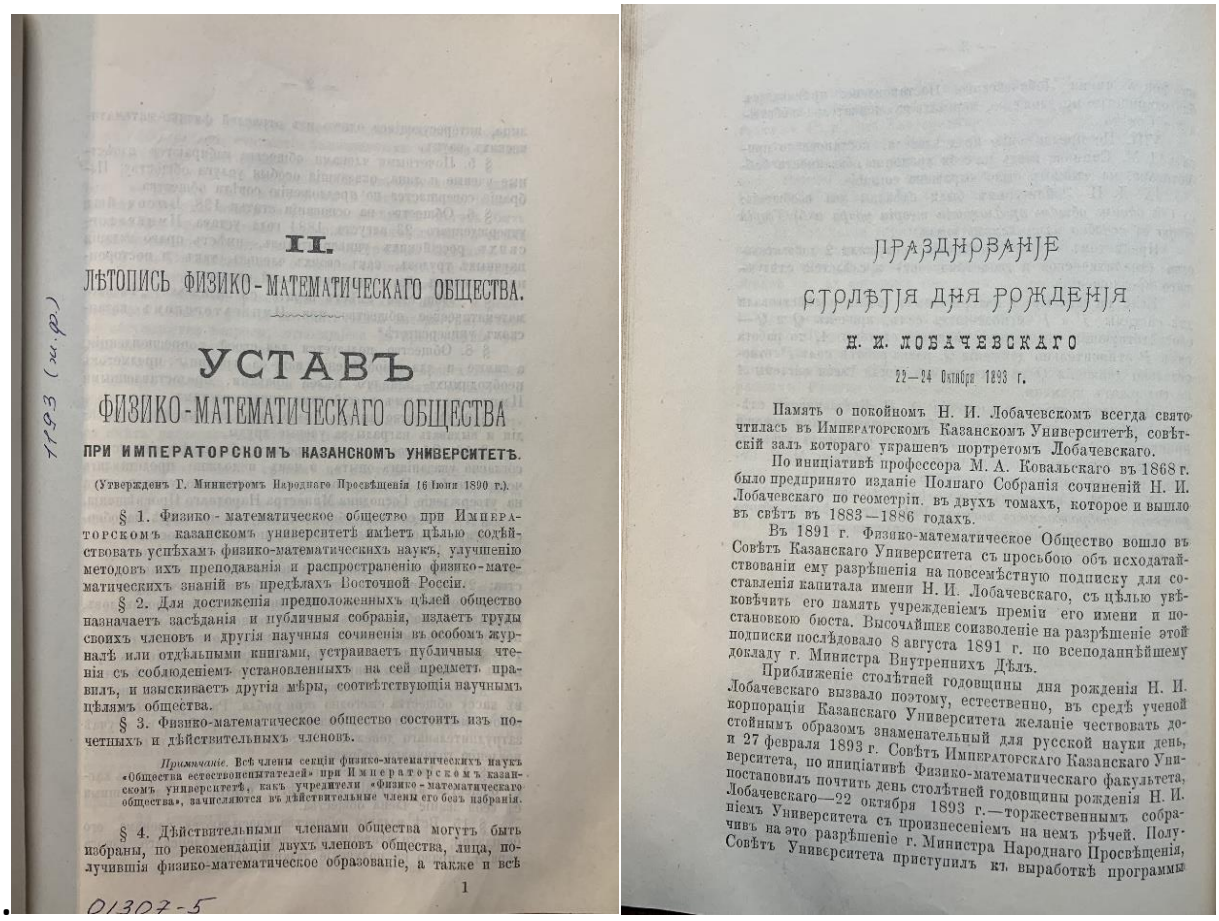
8*

Рис. 7. Перевод «Мемуара» и «Письмо» в «Известиях физико-математического общества при Казанском университете» второй серии.

Алгоритм 3: Экстракция и нормализация метаданных переводных статей

1: **читать** файл номера журнала в формате pdf
2: **загрузить** шаблон, определяющий структурные особенности номера
3: **вычислить** диапазоны страниц статей номера
4: **разделить** файл номера на файлы статей
5: **выделить** первую страницу статьи

6: **осуществить поиск строки** с названием статьи
7: **выделить** название статьи по шрифтовому шаблону
8: **преобразовать** название статьи в метаданные
9: **осуществить поиск строки,** содержащей список авторов
10: **выделить** авторов статьи по соответствующему шаблону
11: **осуществить поиск строки,** содержащей список переводчиков
12: **выделить** переводчиков статьи по соответствующему шаблону
13: **преобразовать** имена переводчиков в именительный падеж
14: **выделить** вторую и третью страницу статьи
15: **найти** номер страницы**, вычислить** номер начальной страницы статьи
16: **выделить** последнюю страницу статьи
17: **осуществить поиск и извлечение** списка литературы
18: **уточнить информацию об авторе** из открытых интернет источников
19: **сформировать** набор метаданных в соответствии схеме нормализации



II.

ЛѢТОПИСЬ ФИЗИКО-МАТЕМАТИЧЕСКАГО ОБЩЕСТВА.

УСТАВЪ
ФИЗИКО-МАТЕМАТИЧЕСКАГО ОБЩЕСТВА
ПРИ ИМПЕРАТОРСКОМЪ КАЗАНСКОМЪ УНИВЕРСИТЕТѢ.

(Утвержденъ Г. Министромъ Народнаго Просвѣщенія 16 Іюня 1890 г.).

§ 1. Физико-математическое общество при Императорскомъ казанскомъ университетѣ имѣетъ цѣлью содѣйствовать успѣхамъ физико-математическихъ наукъ, улучшенію методовъ ихъ преподаванія и распространенію физико-математическихъ знаній въ предѣлахъ Восточной Россіи.

§ 2. Для достиженія предположенныхъ цѣлей общество назначаетъ засѣданія и публичныя собранія, издаетъ труды своихъ членовъ и другія научныя сочиненія въ особомъ журналѣ или отдѣльными книгами, устраиваетъ публичныя чтенія съ соблюденіемъ установленныхъ на сей предметъ правилъ, и изыскиваетъ другія мѣры, соотвѣтствующія научнымъ цѣлямъ общества.

§ 3. Физико-математическое общество состоитъ изъ почетныхъ и дѣйствительныхъ членовъ.

Примѣчаніе. Всѣ члены секціи физико-математическихъ наукъ «Общества естествоиспытателей» при Императорскомъ казанскомъ университетѣ, какъ учредители «Физико-математическаго общества», зачисляются въ дѣйствительные члены его безъ избранія.

§ 4. Дѣйствительными членами общества могутъ быть избраны, по рекомендаціи двухъ членовъ общества, лица, получившія физико-математическое образованіе, а также и всѣ

1

01307-5

ПРАЗДНОВАНІЕ
СТОЛѢТІЯ ДНЯ РОЖДЕНІЯ
Н. И. ЛОБАЧЕВСКАГО
22—24 Октября 1893 г.

Память о покойномъ Н. И. Лобачевскомъ всегда свято чтилась въ Императорскомъ Казанскомъ Университетѣ, совѣтскій залъ котораго украшенъ портретомъ Лобачевскаго.

По иниціативѣ профессора М. А. Ковальскаго въ 1868 г. было предпринято изданіе Полнаго Собранія сочиненій Н. И. Лобачевскаго по геометріи, въ двухъ томахъ, которое и вышло въ свѣтъ въ 1883—1886 годахъ.

Въ 1891 г. Физико-математическое Общество вошло въ Совѣтъ Казанскаго Университета съ просьбою объ исходатайствованіи ему разрѣшенія на повсемѣстную подписку для составленія капитала имени Н. И. Лобачевскаго, съ цѣлью увѣковѣчить его память учрежденіемъ преміи его имени и постановкою бюста. Высочайшее соизволеніе на разрѣшеніе этой подписки послѣдовало 8 августа 1891 г. по всеподданнѣйшему докладу г. Министра Внутреннихъ Дѣлъ.

Приближеніе столѣтней годовщины дня рожденія Н. И. Лобачевскаго вызвало поэтому, естественно, въ средѣ ученой корпораціи Казанскаго Университета желаніе чествовать достойнымъ образомъ знаменательный для русской науки день, и 27 февраля 1893 г. Совѣтъ Императорскаго Казанскаго Университета, по иниціативѣ Физико-математическаго факультета, постановилъ почтить день столѣтней годовщины рожденія Н. И. Лобачевскаго—22 октября 1893 г.—торжественнымъ собраніемъ Университета съ произнесеніемъ на немъ рѣчей. Получивъ на это разрѣшеніе г. Министра Народнаго Просвѣщенія, Совѣтъ Университета приступилъ къ выработкѣ программы

Рис. 8. Разнообразие типов документов в цифровом архиве «Известий физико-математического общества при Казанском университете» второй серии. Приведен

«Устав физико-математического общества при Казанском университете», опубликованный в первом томе журнала в 1891 году и Отчёт о праздновании столетия со дня рождения Н.И. Лобачевского из четвертого тома журнала 1894 года.

На следующем рисунке приведен результат метаописания «Устава …», полученный с помощью алгоритма 4.

```
<article>
    <front>
        <journal-meta>
            <journal-id journal-id-type="pmc">izfmo</journal-id>
            <journal-title-group xml:lang="ru">
                <journal-title>Известия физико-математического общества при Казанском
                Императорском университете</journal-title>
            </journal-title-group>
            <trans-title-group xml:lang="fr">
                <trans-title>Bulletin de la société physico-mathématique de Kasan
                </trans-title>
            </trans-title-group>
            <journal-id journal-id-type="publisher">Kazan</journal-id>
            <publisher>
                <publisher-name>Казань</publisher-name>
            </publisher>
        </journal-meta>
        <article-meta>
            <article-id>2-15-4-1</article-id>
            <title-group>
                <article-title xml:lang="ru">Устав физико-математического общества.
                </article-title>
                <alt-title xml:lang="ru-o">Уставъ физико-математическаго Общества.
                </alt-title>
            </title-group>
            <subj-group>
                <subject>Документ</subject>
            </subj-group>
            <pub-date>
                <year>1891</year>
            </pub-date>
            <volume>1</volume>
            <volume-series>2</volume-series>
            <issue>1</issue>
            <issue-part>2</issue-part>
        </article-meta>
    </front>
</article>
</article>
```

Рис. 9. Фрагмент метаописания архивного документа.

Алгоритм 4: Экстракция и нормализация метаданных документов

1: **читать** файл номера журнала в формате pdf
2: **загрузить** шаблон, определяющий структурные особенности номера
3: **вычислить** диапазоны страниц статей номера
4: **разделить** файл номера на файлы статей
5: **выделить** первую страницу документа
6: **осуществить поиск строки** с названием документа

7: **определить** тип документа
8: **определить** основной язык документа
9: **выделить** название документа по шрифтовому шаблону
10: **перевести** название документа на русский язык (если язык написания – русский дореформенный)
11: **выделить** последнюю страницу статьи
12: **сформировать** набор метаданных в соответствии схеме нормализации

ЗАКЛЮЧЕНИЕ

Представлены решения основных задач, связанных с формированием цифровых математических ретро-коллекций. Приведены алгоритмы создания метаописаний ретро-коллекций, основанные на анализе структуры математических документов, включаемых в них, и применении программных инструментов выделения метаданных. Описаны ретро-коллекции, сформированные с помощью разработанных алгоритмов и включенные в состав цифровой математической библиотеки Lobachevskii-DML. Указаны схемы формирования метаданных и методы нормализации извлеченных метаданных с помощью сервисов, разработанных в рамках фабрики метаданных и в соответствии со схемами и требованиями интегрирующих математических библиотек.

Благодарности

Работа выполнена при финансовой поддержке Российского научного фонда (проект № 21-11-00105).

СПИСОК ЛИТЕРАТУРЫ

1. *Elizarov A.M., Lipachev E.K.* Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // CEUR Workshop Proceedings. 2017. V. 2022. P. 326–333.

2. *Елизаров А.М., Липачёв Е.К.* Семантические методы и инструменты электронной математической библиотеки Lobachevskii-DML // Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (18–23 сентября 2017 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2017. С. 130–136.
URL: http://keldysh.ru/ abrau/ 2017/73.pdf.

3. *Elizarov A.M., Lipachev E.K.* Big Math Methods in Lobachevskii-DML Digital Library // CEUR Workshop Proceedings. 2019. V. 2523. P. 59–72.

4. Developing a 21st Century Global Library for Mathematics Research // Washington: The National Academies Press, 2014. 142 p. https://doi.org/10.17226/18619.

5. *Ion P.* The Effort to Realize a Global Digital Mathematics Library // In: Greuel G.M., Koch T., Paule P., Sommese A. (Eds). Mathematical Software – ICMS 2016. ICMS 2016. Lecture Notes in Computer Science, Springer, Cham, 2016. V. 9725. https://doi.org/10.1007/978-3-319-42432-3_59.

6. *Ion P.D.F., Watt S.M.* The Global Digital Mathematics Library and the International Mathematical Knowledge Trust // ICM 2017: Intelligent Computer Mathematics, 2017. Lecture Notes in Artificial Intelligence. 2017. V. 10383. P. 56–69. https://doi.org/10.1007/978-3-319-62075-6_5.

7. *Bouche T.* Some Thoughts on the Near-Future Digital Mathematics Library. Towards a Digital Mathematics Library. Masaryk University, 2008. P. 3–15. URL: https://eudml.org/doc/221606, last accessed 2020/12/12.

8. *Bouche T.* Digital Mathematics Libraries: The Good, the Bad, the Ugly // Math. Comput. Sci. 2010. V. 3. P. 227–241. https://doi.org/10.1007/s11786-010-0029-2.

9. *Bouche T.* The Digital Mathematics Library as of 2014 // Notices Amer. Math. Soc. 2014. V. 61 (9). P. 1085–1088.

10. EuDML metadata schema specification (v2.0–final), URL: https://initiative.eudml.org/eudml-metadata-schema-specification-v20-final, last accessed 2020/12/12.

11. *Bouche T., Rákosník J.* Report on the EuDML External Cooperation Model // In: Kaiser K., Krantz S.G., Wegner B. (Eds) Topics and Issues in Electronic Publishing, JMM, Special Session. San Diego. 2013. P. 99–108. URL: https://www.emis.de/ proceedings/TIEP2013/07bouche_rakosnik.pdf, last accessed 2020/12/12.

12. *Jost M., Bouche T., Goutorbe C., Jorda J.P.* D3.2: The EuDML metadata schema. Revision: 1.6 as of 15th December 2010. URL: http://www.mathdoc.fr/ publis/d3.2-v1.6.pdf, last accessed 2020/12/12.

13. *Гафурова П.О., Елизаров А.М., Липачёв Е.К., Хамматова Д.М.* Методы

формирования и нормализации метаданных в цифровой математической библиотеке // Научный сервис в сети Интернет: труды XXI Всероссийской научной конференции (23–28 сентября 2019 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2019. С. 234–244. https://doi.org/10.20948/abrau-2019-28.
URL: http://keldysh.ru/abrau/2019/theses/28.pdf, last accessed 2020/12/12.

14. *Gafurova P.O., Elizarov A.M., Lipachev E.K., Khammatova D.M.* Metadata Normalization Methods in the Digital Mathematical Library // CEUR Workshop Proceedings. 2020. V. 2543. P. 136–148.

15. *Zhizhchenko A.B., Izaak A.D.* The information system Math-Net.Ru. Application of contemporary technologies in the scientific work of mathematicians // Russian Math. Surveys. 2007. V. 62 (5). P. 943–966.
http://dx.doi.org/10.1070/RM2007v062n05ABEH004455.

16. *Zhizhchenko A.B., Izaak A.D.* The information system Math-Net.Ru. Current state and prospects. The impact factors of Russian mathematics journals // Russian Math. Surveys. 2009. V. 64 (4). P. 775–784.
http://dx.doi.org/10.1070/RM2009v064n04ABEH004638.

17. *Жижченко А.Б., Изаак А.Д.* Информационная система Math-Net.Ru. Применение современных технологий в научной работе математика // Успехи математических наук. 2007. Т. 62, №5 (377). С. 107–132.
https://doi.org/10.4213/rm8147.
URL: http://www.mathnet.ru/links/c59aff2f134382372f88aa415a76755f/rm8147.pdf.

18. *Жижченко А.Б., Изаак А.Д.* Информационная система Math-Net.Ru. Современное состояние и перспективы развития. Импакт-факторы российских математических журналов // Успехи математических наук. 2009. Т. 64, №4 (388). С. 195–204. https://doi.org/10.4213/rm9312.
URL: http://www.mathnet.ru/links/e27ab619eaefe03fe79d663468ddd3a0/rm9312.pdf

19. *Chebukov D.E., Izaak A.D., Misyurina O.G., Pupyrev Yu.A., Zhizhchenko A.B.* Math-Net.Ru as a Digital Archive of the Russian Mathematical Knowledge from the XIX Century to Today. Intelligent Computer Mathematics // Lecture Notes in Computer Science. 2013. V. 7961. P. 344–348. https://doi.org/ 10.1007/978-3-642-39320-4_26.

20. *Chebukov D.E., Izaak A.D., Misyurina O.G., Pupyrev Yu.A.* Math-Net.Ru video

library: Creating a collection of scientific talks // In: Greuel G.-M. (Ed.) et al., Mathematical software – ICMS 2016. 5th international conference, Berlin, Germany, July 11–14, 2016. Proceedings. Cham: Springer. Lecture Notes in Computer Science. 2016. V. 9725. P. 447–450. https://doi.org/10.1007/978-3-319-42432-3_57.

21. *Гафурова П.О., Елизаров А.М., Липачёв Е.К*. Базовые сервисы цифровой математической библиотеки Lobachevskii-DML // Электронные библиотеки. 2020. Т. 23 (3). С. 336–381. https://doi.org/10.26907/1562-5419-2020-23-3-336-381.

22. *Elizarov A., Lipachev E.* Digital Library Metadata Factories // Proceedings of the International Conference "Internet and Modern Society" (IMS-2020). CEUR Workshop Proceedings. 2021. V. 2813. P. 13–21.

23. *Rocha E.M., Rodrigues J.F.* Disseminating and preserving mathematical knowledge. In: Borwein J.M., Rocha E.M., Rodrigues J.F. (Eds.). Communicating Mathematics in the Digital Era. A K Peters, Ltd., 2008. P. 3–21.

24. *Bouche T.* Toward a Digital Mathematics Library? A French Pedestrian Overview. In: Borwein J.M., Rocha E.M., Rodrigues J.F. (Eds.). Communicating Mathematics in the Digital Era. A K Peters, Ltd., 2008. P. 47–73.

25. *Schonfeld R.* JSTOR a History. Princeton University Press, Princeton, 2003. 448 p.

26. *Burns J., Brenner A., Kiser K., Krot M., Llewellyn C., and Snyder R.* JSTOR – Data for Research // M. Agosti et al. (Eds.): ECDL 2009. Lecture Notes in Computer Science. 2009. V. 5714. P. 416–419.

27. Gallica: the Online Digital Library of the Bibliotheque nationale de France. Review Essay // Nineteenth-Century Music Review. 2014. V. 11 (2). P. 337–347. https://doi.org/10.1017/S1479409814000287.

28. *Bouche T.* The NUMDAM program. MSRI workshop, April 16th 2005, Berkeley, 2005.
URL: https://www.msri.org/specials/dmlp/6-Bouche-numdam.pdf, last accessed 2020/12/12.

29. *Bartošek M., Lhoták M., Rákosník J., Sojka P., and Šárfy M.* The DML-CZ Project: Objectives and First Steps. In: Borwein J.M., Rocha E.M., Rodrigues J.F. (Eds.). Communicating Mathematics in the Digital Era. A K Peters, Ltd., 2008. P. 75–86.

30. *Bartošek M., Rákosník J.* DML-CZ: The Experience of a Medium-Sized Digital

Mathematics Library // Notices of the AMS. 2013. V. 60, No. 8. P. 1028–1033. http://dx.doi.org/10.1090/noti1031.

31. D7.4: Toolset for Image and Text Processing and Metadata Enhancements – Final Release. URL: https://wiki.eudml.eu/mediawiki/eudml/images/D7.4-v1.0.pdf, last accessed 2020/12/12.

32. Journal Article Tag Suite. URL: https://jats.nlm.nih.gov/about.html, last accessed 2020/12/12.

33. *Elizarov A.M., Lipachev E.K.* Methods of Processing Large Collections of Scientific Documents and the Formation of Digital Mathematical Library // CEUR Workshop Proceedings. 2020. V. 2543. P. 354–360.

34. *Nilsson M., Naeve A., Duval E., Johnston P., Massart D.* Harmonization Methodology for Metadata Models. URL: https://hal.archives-ouvertes.fr/hal-00591548, last accessed 2020/12/12.

35. *Elizarov A.M., Lipachev E.K., Haidarov S.M.* Automated Processing Service System of Large Collections of Scientific Documents // CEUR Workshop Proceedings. 2016. V. 1752. P. 58–64.

36. *Elizarov A.M., Khaydarov Sh.M., Lipachev E.R.* Scientific documents ontologies for semantic representation of digital libraries // 2017 Second Russia and Pacific Conference on Computer Technology and Applications (RPC). Vladivostok, Russky Island, Russia 25–29 September, 2017. P. 1–5. https://doi.org/10.1109/RPC.2017.8168064.

37. *Peroni S.* Semantic Web Technologies and Legal Scholarly Publishing, Springer International Publishing, 2014. 304 p. https://doi.org/10.1007/978-3-319-04777-5.

38. *Constantin A., Peroni S., Pettifer S., Shotton D., Vitali F.* The Document Components Ontology (DoCO) // Semantic Web. 2016. V. 7, No. 2. P. 167–181. https://doi.org/10.3233/SW-150177.

39. *Ruiz-Iniesta A., and Corcho O.* A review of ontologies for describing scholarly and scientific documents // CEUR Workshop Proceedings. 2014. V. 1155. P. 1–12. URL: http://ceur-ws.org/Vol-1155/paper-07.pdf, last accessed 2020/12/12.

40. *Kogalovsky M.R., Parinov S.I.* Scholarly Communication in a Semantically Enrichable Research Information System with Embedded Taxonomy of Scientific Relationships

// In: Klinov P., Mouromtsev D. (Eds.) Knowledge Engineering and Semantic Web. Communications in Computer and Information Science, Springer, 2015. V. 518. P. 87–101. https://doi.org/10.1007/978-3-319-24543-0_7.

41. *Биряльцев Е.В., Елизаров А.М., Жильцов Н.Г., Липачёв Е.К., Невзорова О.А., Соловьев В.Д.* Методы анализа семантических данных математических электронных коллекций // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2014. № 4. С. 12–17.

42. *Biryal'tsev E., Elizarov A., Zhil'tsov N., Lipachev E., Nevzorova O., Solov'ev V.* Methods for Analyzing Semantic Data of Electronic Collections in Mathematics // Automatic Documentation and Mathematical Linguistics. 2014. V. 48. No. 2. P. 81–85.

43. *Ronzano F., Saggion H. Dr.* Inventor Framework: Extracting Structured Information from Scientific Publications // In: Japkowicz N., Matwin S. (Eds.) Discovery Science. Lecture Notes in Computer Science, Springer, Cham., 2015. V. 9356. https://doi.org/10.1007/978-3-319-24282-8_18.

44. *Tkaczyk D., Tarnawski B. and Bolikowski Ł.* Structured Affiliations Extraction from Scientific Literature // D-Lib Magazine. 2015. V. 21, No. 11/12. https://doi.org/10.1045/november2015-tkaczyk.

ALGORITHMS FOR FORMATION OF METADATA MATHEMATICAL RETRO COLLECTIONS BASED ON ANALYSIS OF STRUCTURAL FEATURES OF DOCUMENTS

P. O. Gafurova[1] [0000-0002-1544-155X], A. M. Elizarov[2] [0000-0003-2546-6897],
E. K. Lipachev[3] [0000-0001-7789-2332]

[1-3]Kazan Federal University

[1]pogafurova@gmail.com, [2] amelizarov@gmail.com, [3] elipachev@gmail.com

Abstract

The solutions of the main problems associated with the formation of digital mathematical collections from documents published in the pre-digital period are presented – such collections are designated in the work as retro collections. Algorithms for creating a meta description of retro collections based on the analysis of the structure of mathematical documents and the use of software tools for extracting metadata are given. The description of retro-collections formed using the developed algorithms and included in the metadata factory of the digital mathematical library Lobachevskii-DML is given. The schemes for the formation of metadata and methods for normalizing the extracted metadata in accordance with the schemes and requirements of the integrating mathematical libraries are indicated.

Keywords*: Lobachevskii-DML, metadata factory, metadata management services, archive collections.*

REFERENCES

1. *Elizarov A.M., Lipachev E.K.* Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // CEUR Workshop Proceedings. 2017. V. 2022. P. 326–333.

2. *Elizarov A.M., Lipachev E.K.* Semanticheskie metody i instrumenty ehlektronnoj matematicheskoj biblioteki Lobachevskii-DML // Nauchnyj servis v seti Internet: trudy XIX Vserossijskoj nauchnoj konferencii (18–23 sentyabrya 2017 g., g. Novorossijsk). M.: IPM im. M.V. Keldysha, 2017. S. 130–136.

https://doi.org/10.20948/abrau-2017-73.
URL: http://keldysh.ru/ abrau/ 2017/73.pdf.

3. *Elizarov A.M., Lipachev E.K.* Big Math Methods in Lobachevskii-DML Digital Library // CEUR Workshop Proceedings. 2019. V. 2523. P. 59–72.

4. Developing a 21st Century Global Library for Mathematics Research // Washington: The National Academies Press, 2014. 142 p. http://dx.doi.org/10.17226/18619.

5. *Ion P.* The Effort to Realize a Global Digital Mathematics Library // In: Greuel G.M., Koch T., Paule P., Sommese A. (Eds). Mathematical Software – ICMS 2016. ICMS 2016. Lecture Notes in Computer Science, Springer, Cham, 2016. V. 9725. https://doi.org/10.1007/978-3-319-42432-3_59.

6. *Ion P.D.F., Watt S.M.* The Global Digital Mathematics Library and the International Mathematical Knowledge Trust // ICM 2017: Intelligent Computer Mathematics, 2017. Lecture Notes in Artificial Intelligence. 2017. V. 10383. P. 56–69. https://doi.org/10.1007/978-3-319-62075-6_5.

7. *Bouche T.* Some Thoughts on the Near-Future Digital Mathematics Library. Towards a Digital Mathematics Library. Masaryk University, 2008. P. 3–15.
URL: https://eudml.org/doc/221606, last accessed 2020/12/12.

8. *Bouche T.* Digital Mathematics Libraries: The Good, the Bad, the Ugly // Math. Comput. Sci. 2010. V. 3. P. 227–241. https://doi.org/10.1007/s11786-010-0029-2.

9. *Bouche T.* The Digital Mathematics Library as of 2014 // Notices Amer. Math. Soc. 2014. V. 61 (9). P. 1085–1088.

10. EuDML metadata schema specification (v2.0–final),
URL: https://initiative.eudml.org/eudml-metadata-schema-specification-v20-final, last accessed 2020/12/12.

11. *Bouche T., Rákosník J.* Report on the EuDML External Cooperation Model // In: Kaiser K., Krantz S.G., Wegner B. (Eds) Topics and Issues in Electronic Publishing, JMM, Special Session. San Diego. 2013. P. 99–108.
URL: https://www.emis.de/ proceedings/TIEP2013/07bouche_rakosnik.pdf, last accessed 2020/12/12.

12. *Jost M., Bouche T., Goutorbe C., Jorda J.P.* D3.2: The EuDML metadata schema. Revision: 1.6 as of 15th December 2010.

URL: http://www.mathdoc.fr/publis/d3.2-v1.6.pdf, last accessed 2020/12/12.

13. *Gafurova P.O., Elizarov A.M., Lipachev E.K., Khammatova D.M.* Methods of Formation and Normalization of Metadata in the Digital Mathematical Library // Nauchnyj servis v seti Internet: trudy XXI Vserossijskoj nauchnoj konferencii (23–28 sentyabrya 2019 g., g. Novorossijsk). M.: IPM im. M.V. Keldysha, 2019. S. 234–244.
https://doi.org/10.20948/abrau-2019-28.
URL: http://keldysh.ru/abrau/2019/theses/28.pdf, last accessed 2020/12/12.

14. *Gafurova P.O., Elizarov A.M., Lipachev E.K., Khammatova D.M.* Metadata Normalization Methods in the Digital Mathematical Library // CEUR Workshop Proceedings. 2020. V. 2543. P. 136–148.

15. *Zhizhchenko A.B., Izaak A.D.* The information system Math-Net.Ru. Application of contemporary technologies in the scientific work of mathematicians // Russian Math. Surveys. 2007. V. 62 (5). P. 943–966.
http://dx.doi.org/10.1070/ RM2007v062n05ABEH004455.

16. *Zhizhchenko A.B., Izaak A.D.* The information system Math-Net.Ru. Current state and prospects. The impact factors of Russian mathematics journals // Russian Math. Surveys. 2009. V. 64 (4). P. 775–784.
http://dx.doi.org/10.1070/RM2009v064n04ABEH004638.

17. *Zhizhchenko A.B., Izaak A.D.* Informacionnaya sistema Math-Net.Ru. Primenenie sovremennykh tekhnologij v nauchnoj rabote matematika // Uspekhi matematicheskikh nauk. 2007. T. 62, №5 (377). S. 107–132.
https://doi.org/10.4213/rm8147.
URL: http://www.mathnet.ru/links/ c59aff2f134382372f88aa415a76755f/rm8147.pdf.

18. *Zhizhchenko A.B., Izaak A.D.* Informacionnaya sistema Math-Net.Ru. Sovremennoe sostoyanie i perspektivy razvitiya. Impakt-faktory rossijskikh matematicheskikh zhurnalov // Uspekhi matematicheskikh nauk. 2009. T. 64, №4 (388). S. 195–204. https://doi.org/10.4213/rm9312;
URL: http://www.mathnet.ru/links/e27ab619eaefe03fe79d663468ddd3a0/rm9312.pdf

19. *Chebukov D.E., Izaak A.D., Misyurina O.G., Pupyrev Yu.A., Zhizhchenko A.B.* Math-Net.Ru as a Digital Archive of the Russian Mathematical Knowledge from the XIX Century to Today. Intelligent Computer Mathematics // Lecture Notes in Computer Science.

2013. V. 7961. P. 344–348. https://doi.org/ 10.1007/978-3-642-39320-4_26.

20. *Chebukov D.E., Izaak A.D., Misyurina O.G., Pupyrev Yu.A.* Math-Net.Ru video library: Creating a collection of scientific talks // In: Greuel G.-M. (Ed.) et al., Mathematical software – ICMS 2016. 5th International Conference, Berlin, Germany, July 11–14, 2016. Proceedings. Cham: Springer. Lecture Notes in Computer Science. 2016. V. 9725. P. 447–450. https://doi.org/10.1007/978-3-319-42432-3_57.

21. *Гафурова П.О., Елизаров А.М., Липачёв Е.К*. Базовые сервисы цифровой математической библиотеки Lobachevskii-DML // Электронные библиотеки. 2020. Т. 23 (3). С. 336–381. https://doi.org/10.26907/1562-5419-2020-23-3-336-381.

22. *Elizarov A., Lipachev E.* Digital Library Metadata Factories // Proceedings of the International Conference "Internet and Modern Society" (IMS-2020). CEUR Workshop Proceedings. 2021. V. 2813. P. 13–21.

23. *Rocha E.M., Rodrigues J.F.* Disseminating and preserving mathematical knowledge. In: Borwein J.M., Rocha E.M., Rodrigues J.F. (Eds.). Communicating Mathematics in the Digital Era. A K Peters, Ltd., 2008. P. 3–21.

24. *Bouche T.* Toward a Digital Mathematics Library? A French Pedestrian Overview. In: Borwein J.M., Rocha E.M., Rodrigues J.F. (Eds.). Communicating Mathematics in the Digital Era. A K Peters, Ltd., 2008. P. 47–73.

25. *Schonfeld R.* JSTOR a History. Princeton University Press, Princeton, 2003. 448 p.

26. *Burns J., Brenner A., Kiser K., Krot M., Llewellyn C., Snyder R.* JSTOR – Data for Research // M. Agosti et al. (Eds.): ECDL 2009. Lecture Notes in Computer Science. 2009. V. 5714. P. 416–419.

27. Gallica: the Online Digital Library of the Bibliothe`que nationale de France. Review Essay // Nineteenth-Century Music Review. 2014. V. 11 (2). P. 337–347. https://doi.org/10.1017/S1479409814000287.

28. *Bouche T.* The NUMDAM program. MSRI workshop, April 16th 2005, Berkeley, 2005. URL: https://www.msri.org/specials/dmlp/6-Bouche-numdam.pdf, last accessed 2020/12/12.

29. *Bartošek M., Lhoták M., Rákosník J., Sojka P., Šárfy M.* The DML-CZ Project: Objectives and First Steps. In: Borwein J.M., Rocha E.M., Rodrigues J.F. (Eds.). Communicating Mathematics in the Digital Era. A K Peters, Ltd., 2008. P. 75–86.

30. *Bartošek M., Rákosník J.* DML-CZ: The Experience of a Medium-Sized Digital Mathematics Library // Notices of the AMS. 2013. V. 60, No. 8. P. 1028–1033. http://dx.doi.org/10.1090/noti1031.

31. D7.4: Toolset for Image and Text Processing and Metadata Enhancements – Final Release. URL: https://wiki.eudml.eu/mediawiki/eudml/images/D7.4-v1.0.pdf, last accessed 2020/12/12.

32. Journal Article Tag Suite. https://jats.nlm.nih.gov/about.html, last accessed 2020/12/12.

33. *Elizarov A.M., Lipachev E.K.* Methods of Processing Large Collections of Scientific Documents and the Formation of Digital Mathematical Library // CEUR Workshop Proceedings. 2020. V. 2543. P. 354–360.

34. *Nilsson M., Naeve A., Duval E., Johnston P., Massart D.* Harmonization Methodology for Metadata Models. https://hal.archives-ouvertes.fr/hal-00591548, last accessed 2020/12/12.

35. *Elizarov A.M., Lipachev E.K., Haidarov S.M.* Automated Processing Service System of Large Collections of Scientific Documents // CEUR Workshop Proceedings. 2016. V. 1752. P. 58–64.

36. *Elizarov A.M., Khaydarov Sh.M., Lipachev E.K.* Scientific documents ontologies for semantic representation of digital libraries // 2017 Second Russia and Pacific Conference on Computer Technology and Applications (RPC). Vladivostok, Russky Island, Russia 25-29 September, 2017. P. 1–5. https://doi.org/10.1109/RPC.2017.8168064.

37. *Peroni S.* Semantic Web Technologies and Legal Scholarly Publishing, Springer International Publishing, 2014. 304 p. https://doi.org/10.1007/978-3-319-04777-5.

38. *Constantin A., Peroni S., Pettifer S., Shotton D., Vitali F.* The Document Components Ontology (DoCO) // Semantic Web. 2016. V. 7, No. 2. P. 167–181. https://doi.org/10.3233/SW-150177.

39. *Ruiz-Iniesta A., Corcho O.* A review of ontologies for describing scholarly and scientific documents // CEUR Workshop Proceedings. 2014. V. 1155. P. 1–12. URL: http://ceur-ws.org/Vol-1155/paper-07.pdf, last accessed 2020/12/12.

40. *Kogalovsky M.R., Parinov S.I.* Scholarly Communication in a Semantically Enrichable Research Information System with Embedded Taxonomy of Scientific Relationships

// In: Klinov P., Mouromtsev D. (Eds.) Knowledge Engineering and Semantic Web. Communications in Computer and Information Science, Springer, 2015. V. 518. P. 87–101. https://doi.org/10.1007/978-3-319-24543-0_7.

41. *Biryal'cev E.V., Elizarov A.M., Zhil'cov N.G., Lipachev E.K., Nevzorova O.A., Solov'ev V.D.* Metody analiza semanticheskikh dannykh matematicheskikh ehlek-tronnykh kollekcij // Nauchno-tekhnicheskaya informaciya. Seriya 2: Informacionnye processy i sistemy. 2014. № 4. S. 12–17.

42. *Biryal'tsev E., Elizarov A., Zhil'tsov N., Lipachev E., Nevzorova O., Solov'ev V.* Methods for Analyzing Semantic Data of Electronic Collections in Mathematics // Automatic Documentation and Mathematical Linguistics. 2014. V. 48. No. 2. P. 81–85.

43. *Ronzano F., Saggion H.Dr.* Inventor Framework: Extracting Structured Information from Scientific Publications // In: Japkowicz N., Matwin S. (Eds.) Discovery Science. Lecture Notes in Computer Science, Springer, Cham, 2015. V. 9356.
https://doi.org/10.1007/978-3-319-24282-8_18.

44. *Tkaczyk D., Tarnawski B., Bolikowski Ł.* Structured Affiliations Extraction from Scientific Literature // D-Lib Magazine. 2015. V. 21, No. 11/12.
https://doi.org/10.1045/november2015-tkaczyk.

СВЕДЕНИЯ ОБ АВТОРАХ

ГАФУРОВА Полина Олеговна – магистр математики, аспирант Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Polina GAFUROVA – Magister of Mathematics, Kazan (Volga Region) Federal University.

Current scientific interests: data mining, recommender systems, cloud computing, knowledge extraction technologies.

email: pogafurova@gmail.com; ORCID: 0000-0002-1544-155X

ЕЛИЗАРОВ Александр Михайлович – доктор физико-математических наук, профессор, заслуженный деятель науки Республики Татарстан, профессор кафедры программной инженерии Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Alexander ELIZAROV – Doctor of Physics and Mathematics, Professor, Honoured Worker of Science of the Republic of Tatarstan, Kazan Federal University.

Current scientific interests: data mining, recommender systems, cloud computing, knowledge extraction technologies.

email: amelizarov@gmail.com; ORCID: 0000-0003-2546-6897

ЛИПАЧЁВ Евгений Константинович – кандидат физико-математических наук, доцент кафедры Интеллектуальных технологий поиска Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Evgeny Konstantinovich LIPACHEV – Candidate of Physics and Mathematics, Associate Professor, Kazan Federal University.

Current scientific interests: data mining, recommender systems, cloud computing, knowledge extraction technologies.

email: elipachev@gmail.com; ORCID: 0000-0001-7789-2332

Материал поступил в редакцию 21 ноября 2020 года

Переработанная версия – 16 апреля 2021 года

УДК 004.4+004.6+004.8+004.9

ПРИМЕНЕНИЕ МАШИННОГО ОБУЧЕНИЯ К ЗАДАЧЕ ГЕНЕРАЦИИ ПОИСКОВЫХ ЗАПРОСОВ

А. М. Гусенков[1, [0000-0003-4019-7322]], А. Р. Ситтикова[2, [0000-0002-9539-764X]]

[1,2]*Казанский* (Приволжский) федеральный университет, Казань, Россия

[1]gusenkov.a.m@gmail.com, [2]sitti.alina@mail.ru

Аннотация

Исследованы две модификации рекуррентных нейронных сетей: сети с долгой краткосрочной памятью и сети с управляемым рекуррентным блоком с добавлением механизма внимания к обеим сетям, а также модель Transformer в задаче генерации запросов к поисковым системам. В качестве модели Transformer использована модель GPT-2 от OpenAI, которая обучалась на запросах пользователей. Проведен латентно-семантический анализ для определения семантических сходств между корпусом пользовательских запросов и запросов, генерируемых нейронными сетями. Для проведения анализа корпус был переведен в формат bag of words, к нему применена модель TFIDF, проведено сингулярное разложение. Семантическое сходство вычислялось на основе косинусной меры. Также для более полной оценки применимости моделей к задаче был проведен экспертный анализ для оценки связности слов в искусственно созданных запросах.

Ключевые слова: *обработка естественного языка, генерация естественного языка, машинное обучение, нейронные сети.*

ВВЕДЕНИЕ

Генерация естественного языка – это процесс создания осмысленных фраз и предложений в форме естественного языка. Среди основных применяемых подходов можно выделить два алгоритма создания текстов: методы на основе правил и методы на основе машинного обучения. Первый подход позволяет добиться высокого качества текстов, но требует знания правил языка и времени для разработки [1], в то время как второй подход зависит только от данных для обучения, но часто допускает грамматические и семантические ошибки в создаваемых текстах [2].

В настоящее время активно исследуется метод генерации текстов с помощью нейронных сетей; один из самых популярных алгоритмов – рекуррентные нейронные сети [3]. Вторая ведущая архитектура – модель Transformer [3]. Эти архитектуры рассматривались в решении задачи генерации поисковых запросов.

Цель данной статьи – изучить вышеупомянутые архитектуры, проанализировать их качество и применимость к этой задаче. Использование автоматически генерируемых запросов для поисковых систем актуально, поскольку большинство компаний не выдает поисковые запросы бесплатно, а поисковая система в процессе разработки должна быть протестирована. Также полученные запросы можно использовать для повышения эффективности и оптимизации поисковой системы.

Были использованы поисковые запросы от пользователей AOL (America Online), которые были анонимно размещены в интернете в 2006 г. Хотя названная компания не идентифицировала своих пользователей, личная информация присутствовала во многих запросах [4], чего компании сейчас пытаются избежать. Были предложены алгоритмы, помогающие сохранить анонимность пользователей, но существует вопрос о том, имеют ли данные, которые можно безопасно публиковать, практическую пользу. Для решения этой проблемы предлагается использовать автоматически генерируемые запросы.

ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

Алгоритмы генерации текста на естественном языке активно изучаются и используются во многих программных системах, поэтому на данный момент ведется большое количество исследований в этой области.

Один из первых применяемых подходов – это система шаблонов для заполнения пробелов. Она используется в текстах, которые имеют предопределенную структуру, и, если необходимо заполнить небольшой объем данных, этот подход может автоматически заполнять пробелы данными, полученными из электронных таблиц, баз данных и т. д. Примером такого подхода является Microsoft Word mailmerge [5].

Вторым шагом было добавление к первому подходу языков программирования общего назначения, которые поддерживают сложные условные выражения, циклы и т. д. Этот подход более эффективный и полезный, но отсутствие языковых возможностей затрудняет создание систем, которые могут генерировать качественные тексты.

Следующим шагом в развитии систем на основе шаблонов является добавление грамматических функций на уровне слов, связанных с морфологией и правописанием. Такие функции значительно упрощают создание грамматически правильных текстов. Далее системы динамически создают предложения из представлений значений, которые они должны передать. Это означает, что системы могут обрабатывать необычные случаи без необходимости явного написания кода для каждого случая и значительно лучше генерируют высококачественного тексты на «микроуровне». Наконец, на следующем этапе развития системы могут генерировать хорошо структурированные документы, актуальные для пользователей. Например, текст, который должен быть убедительным, может быть основан на моделях аргументации и изменения поведения [5].

После перехода от шаблонов к генерации динамического текста потребовалось много времени, чтобы добиться удовлетворительных результатов. Если рассматривать создание текстов на естественном языке как подраздел обработки естественного языка, то существует ряд наиболее развитых алгоритмов: цепи Маркова [6], рекуррентные нейронные сети, сети с долгой краткосрочной памятью и модель Transformer. Существуют инструменты для генерации текста, основанные на этих методах, например, коммерческие Arria NLG PLC, AX Semantics, Yseop и другие, а также программы с открытым исходным кодом Simplenlg, GPT, GPT-2, BERT, XLNet.

Кроме того, в настоящее время исследуется использование генеративно-состязательных сетей для генерации текста, поскольку они показывают отличные результаты в задаче генерации изображений [7].

СБОР ДАННЫХ

В качестве обучающих данных для нейронных сетей были выбраны пользовательские запросы на английском языке из поисковой системы AOL 2006 года. Исследователи стараются избегать использования этих данных в своих работах,

так как они могут считаться разоблачающими, но в данной статье используются только тексты самих запросов, без идентификаторов пользователей и сайтов, на которые они перешли, то есть без использования личной информации. Исходные данные представлены в виде, показанном на рис. 1.

```
Query   QueryTime       ItemRank        ClickURL
carbol tunnel   2006-03-01 01:01:21
how to install a glue down floor        2006-03-01 07:13:45     2       http://doityourself.com
how to install a glue down floor        2006-03-01 07:13:45     8       http://www.homerenovationguide.com
how to install a glue down floor        2006-03-01 07:13:45     9       http://www.hardwoodinstaller.com
how to install a glue down floor        2006-03-01 07:35:01     20      http://www.ehow.com
how to install a glue down floor        2006-03-01 07:43:50     26      http://www.hoskinghardwood.com
mapquest        2006-03-01 19:40:11     1       http://www.mapquest.com
indian projectile points        2006-03-02 21:12:10
indian projectile points        2006-03-02 21:13:02
indian projectile points        2006-03-02 21:13:03     1       http://www.utexas.edu
indian projectile points        2006-03-02 21:13:03     6       http://www.iath.virginia.edu
indian projectile points        2006-03-02 21:22:40
indian projectile points        2006-03-02 21:22:42
indian projectile points        2006-03-02 21:22:46     16      http://www.mnsu.edu
indian projectile points        2006-03-02 21:22:46     18      http://www.madison.k12.wi.us
```

Рис. 1. Исходные данные для обучения.

Запросы длиной более 32 слов и ошибочные запросы, не содержащие информации, были удалены из корпуса. Повторяющиеся запросы и запросы, содержащие названия веб-сайтов, также были удалены, поскольку они не являются примерами естественного языка. Всего были случайным образом выбраны 100 тыс. запросов для обучения. На рис. 2 показаны примеры данных после предварительной обработки.

```
i can love you like that
breyer traditional horse models
waffle irons
home made structures
dominion a prequel to the exorcist
what is a liability
capstone turbine
danville ill
statetax
old club men ties
piciculture
pagan jewelry
unicoi county memorial hospital
```

Рис. 2. Данные после предварительной обработки.

Запросы были разделены на символы-токены, каждому символу было присвоено натуральное число, весь корпус был закодирован с помощью этого словаря.

РЕКУРРЕНТНЫЕ СЕТИ

Рекуррентные нейронные сети (Recurrent Neural Networks, RNN) – это семейство нейронных сетей, в которых связи между элементами образуют направленную последовательность [8]. Они могут использовать свою внутреннюю память для обработки последовательностей произвольной длины, а также хорошо распознают зависимости между токенами. Однако рекуррентные сети учатся медленно, и их способность запоминать длинные зависимости ограничена из-за проблемы затухания градиентов [9].

Были реализованы два типа рекуррентных сетей, которые чаще всего используются в задаче генерации текстов на естественном языке – сети с долгой краткосрочной памятью (Long Short-Term Memory Network) [10] и сети с управляемым рекуррентным блоком (Gated Recurrent Unit) [11]. Исследования показали, что эти типы сетей имеют сопоставимую точность, а в зависимости от решаемой задачи одна сеть может быть точнее другой.

СЕТИ С ДОЛГОЙ КРАТКОСРОЧНОЙ ПАМЯТЬЮ

Сеть с долгой краткосрочной памятью (Long Short-Term Memory Networks, LSTM) – это система глубокого обучения, которая позволяет избежать проблемы затухания и взрывного роста градиентов [10]. Сети LSTM могут запоминать значительно более длинные последовательности символов. Они используют вентили, которые являются внутренними механизмами, которые могут контролировать информационный поток. На рис. 3 представлен общий вид ячейки LSTM.

В каждой ячейке сети есть 3 вентиля: входной, выходной и вентиль забывания. Вектор вентиля забывания вычисляется по формуле

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f),$$

где x_t – входной вектор, h_{t-1} – выходной вектор предыдущей ячейки, σ – сигмоидальная функция, W_f, U_f, b_f – матрицы весов и вектор смещения.

Далее входной вентиль обновляет состояние ячейки по следующим формулам:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \qquad \hat{c}_t = tanh(W_c x_t + U_c h_{t-1} + b_c).$$

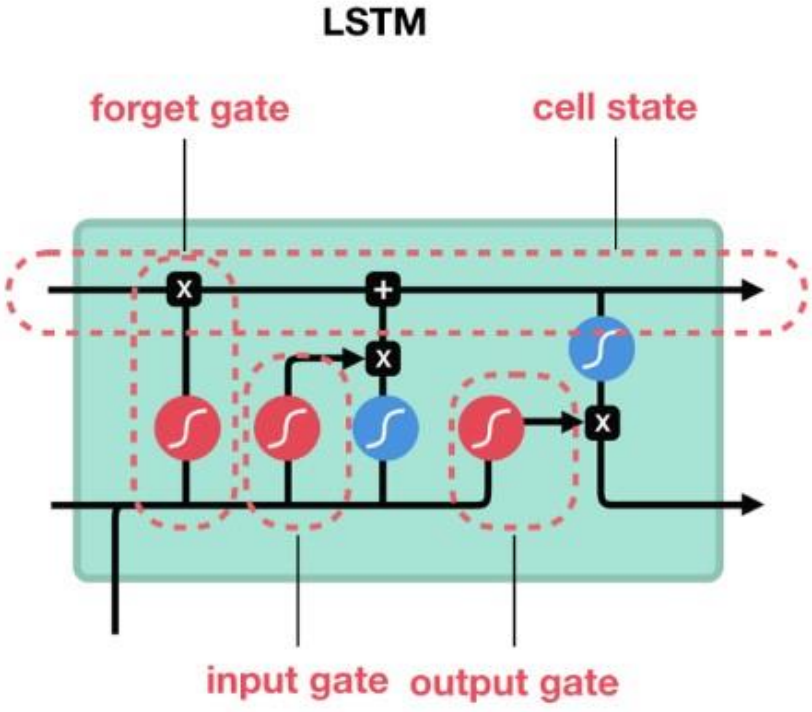


Рис. 3. Ячейка LSTM.

Затем вычисляется новое значение состояния ячейки:

$$c_t = f_t \circ c_{t-1} + i_t \circ \hat{c}_t,$$

где c_{t-1} – состояние предыдущей ячейки. Наконец, выходной вентиль решает, каким должно быть следующее скрытое состояние по формулам:

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o),\ h_t = o_t \circ \tanh(c_t).$$

Результаты передаются в следующую ячейку.

СЕТИ С УПРАВЛЯЕМЫМ РЕКУРРЕНТНЫМ БЛОКОМ

Вторая реализованная модель – это сеть с управляемым рекуррентным блоком (Gated Recurrent Unit, GRU) – новое поколение рекуррентных нейронных сетей, похожих на сети с долгой краткосрочной памятью [11]. Однако по сравнению с LSTM этот тип сетей имеет меньше параметров, и поэтому эти модели обучаются быстрее. У GRU всего два вентиля: обновления и сброса. На рис. 4 представлен общий вид ячейки сети GRU.

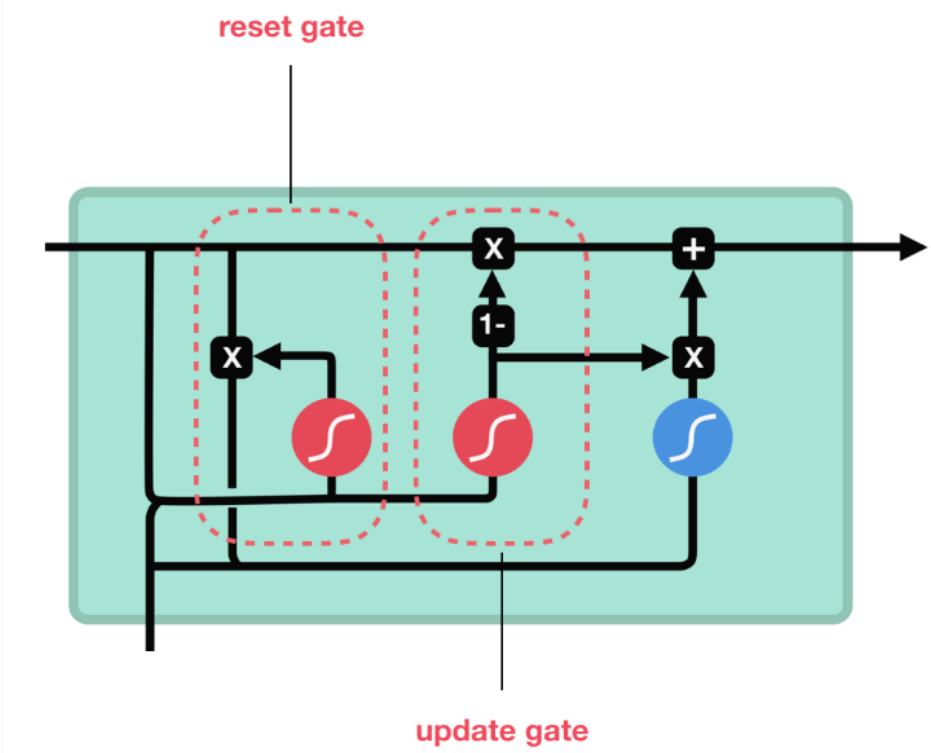


Рис. 4. Ячейка GRU

Вентиль обновления действует подобно входному вентилю и вентилю забывания в LSTM и вычисляется по формуле

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z).$$

Вентиль сброса рассчитывается по формуле

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r).$$

Выходной вектор ячейки GRU определяется следующим образом:

$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \sigma(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h).$$

МЕХАНИЗМ ВНИМАНИЯ В РЕКУРРЕНТНЫХ НЕЙРОННЫХ СЕТЯХ

Механизм внимания – это метод, используемый в нейронных сетях для выявления зависимостей между частями входных и выходных данных [12]. Механизм внимания позволяет модели определять важность каждого слова для задачи прогнозирования путем их взвешивания при создании представления текста. Был использован следующий подход с одним параметром на входной канал [13]:

$$e_t = h_t w_a,\ a_t = \frac{\exp(e_t)}{\sum_{i=1}^{T} \exp(e_i)},\ v = \sum_{i=1}^{T} a_i h_i\ .$$

Здесь h_t – представление слова в момент времени t, w_a – матрица весов для слоя внимания, a_t – оценки внимания для каждого момента времени, а v – вектор представления текста.

РЕАЛИЗАЦИЯ РЕКУРРЕНТНЫХ СЕТЕЙ

Все нейронные сети были реализованы на Python 3.7 в пакете Google Collab [14], поскольку он позволяет использовать графические процессоры, что значительно сокращает время обучения моделей. Для реализации нейронных сетей мы выбрали библиотеку Keras [15], которая представляет собой высокоуровневую надстройку над TensorFlow. Эта библиотека значительно упрощает разработку нейронных сетей, так как в ней уже есть готовые реализации основных слоев, функции активации и потери. Был использован оптимизатор Adam (Adaptive Moment Estimation [16]) – алгоритм, в котором скорость обучения регулируется для каждого параметра. Также использована функция Learning Rate Scheduler [17] в качестве callback, которая позволяет вычислять коэффициент скорости обучения с помощью определенной функции. Примерная архитектура модели представлена на рис. 5.

Предварительно обработанные данные были разделены на обучающую и валидационную выборки, которые составили 80% и 20% корпуса соответственно. Обучающие данные передаются на вход слою Embedding, который преобразует числа в векторы, отражающие соответствия между последовательностями символов и проекции этих последовательностей. Полученные представления передаются на вход первому слою LSTM (GRU), его выходные данные передаются второму слою LSTM (GRU), и тем же образом – третьему. Затем выходные данные из слоя Embedding и этих трех слоев объединяются и передаются слою Attention-WeightedAverage. Вектор представления, полученный из слоя внимания, представляет собой высокоуровневое кодирование всего текста, которое используется в качестве входных данных для конечного полносвязного слоя с активацией Softmax для классификации [13].

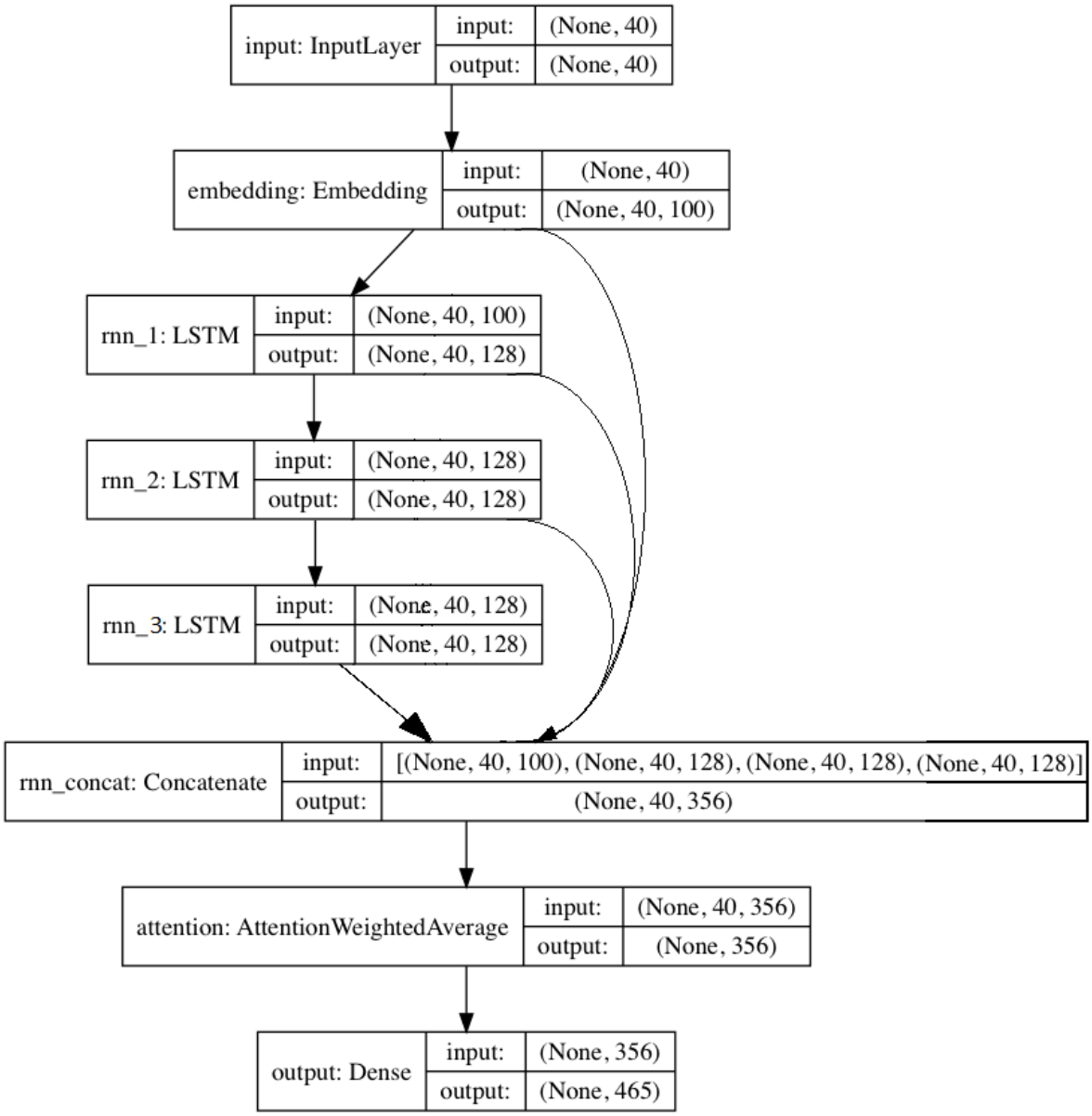


Рис. 5. Архитектура модели.

Для проверки, насколько хорошо или плохо модель обучилась за конкретную эпоху, вычисляется функция потерь Categorical Cross-Entropy по формуле

$$L(y, \hat{y}) = -\sum_{j=0}^{M}\sum_{i=0}^{N}\left(y_{ij}\log\left(\hat{y}_{ij}\right)\right)$$

где $\hat{y}$ – предсказанные значения.

Были проведены эксперименты с изменением количества слоев LSTM (GRU) в модель (2 и 3 слоя), а также с добавлением слоя Dropout после слоя Embedding,

который случайным образом исключает заданное количество нейронов, чтобы предотвратить переобучение сети и лучше обобщить модель. Сети с 3 рекуррентными слоями и Dropout показали лучший результат.

Также были обучены двунаправленные модели этих сетей. Двунаправленная рекуррентная нейронная сеть – это модель, предложенная в 1997 году Майком Шустером и Кулдип Паливал [18], которая позволяет рассматривать контекст слова не только слева от него, но и справа от последовательности. Общий вид двунаправленных нейронных сетей представлен на рис. 6.

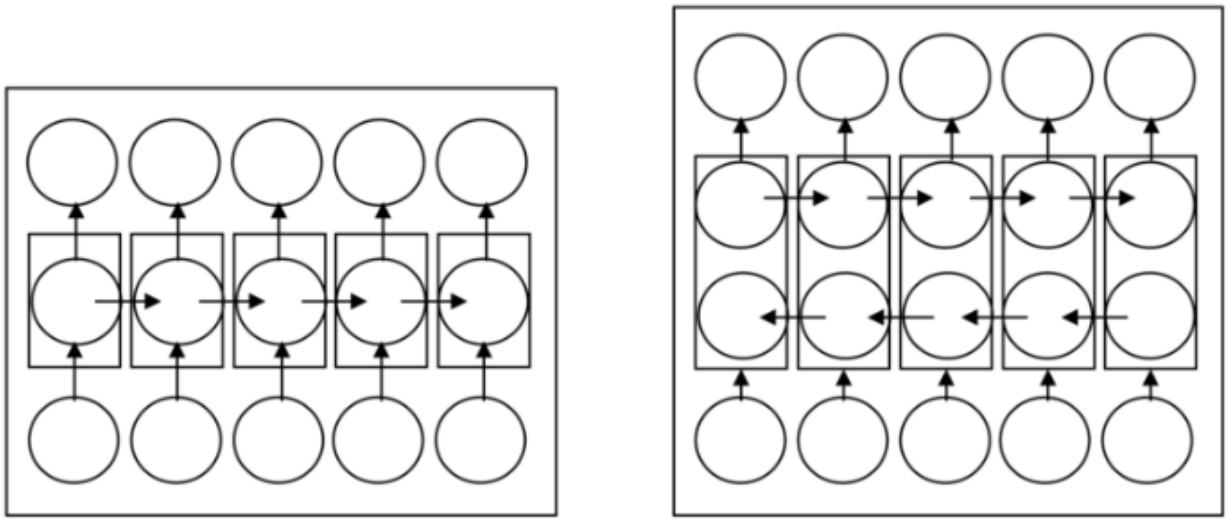


Рис. 6. Общий вид: а) Однонаправленная нейронная сеть;
б) Двунаправленная нейронная сеть

В случае задачи генерации поисковых запросов двунаправленная модель показала себя лучше однонаправленной; полученные значения функции потерь после обучения модели в течение 30 эпох представлены в таблице 1.

Таблица 1. Значения функции потерь

	LSTM	GRU	Bi-LSTM	Bi-GRU
Loss	1,48	1,58	1,30	1,37
Validation Loss	1,6	1,62	1,56	1,57

Значение функции потерь уменьшилось на обучающих данных, однако на валидационных данных улучшение было менее значительным, что позволяет предположить, что двунаправленная модель в этом задаче не так хорошо обучается, а скорее «запоминает» последовательности символов.

С помощью реализованной модели были сгенерированы запросы с разной «температурой». Это параметр, который влияет на шанс выбора маловероятного символа.

ТРАНСФОРМЕР

Transformer – это модель глубокого обучения, которая была представлена в 2017 году [19]. Общий вид его архитектуры показан на рис. 7.

Трансформеры состоят из стэков равного количества энкодеров и декодеров. Энкодеры обрабатывают входные последовательности и кодируют данные для отражения информации о них и их признаках. Декодеры делают обратное, они обрабатывают полученную от энкодера информацию и генерируют выходные последовательности. Все энкодеры имеют одинаковую структуру и состоят из двух слоёв: внутреннее внимание (Self-Attention) и нейронная сеть с прямой связью (feed-forward neural network). Входная последовательность, поступающая в энкодер, сначала проходит через слой внутреннего внимания, помогающий энкодеру посмотреть на другие слова во входном предложении при кодировании конкретного слова. Выходные данные этого слоя отправляются в нейронную сеть с прямой связью. Такая же сеть независимо применяется к каждому слову. Декодер также содержит два этих слоя, но между ними есть дополнительный слой внимания, который позволяет декодеру определить релевантные части входного предложения.

Внутреннее внимание позволяет модели видеть зависимости между обрабатываемым словом и другими словами во входной последовательности, которые помогают лучше закодировать слово.

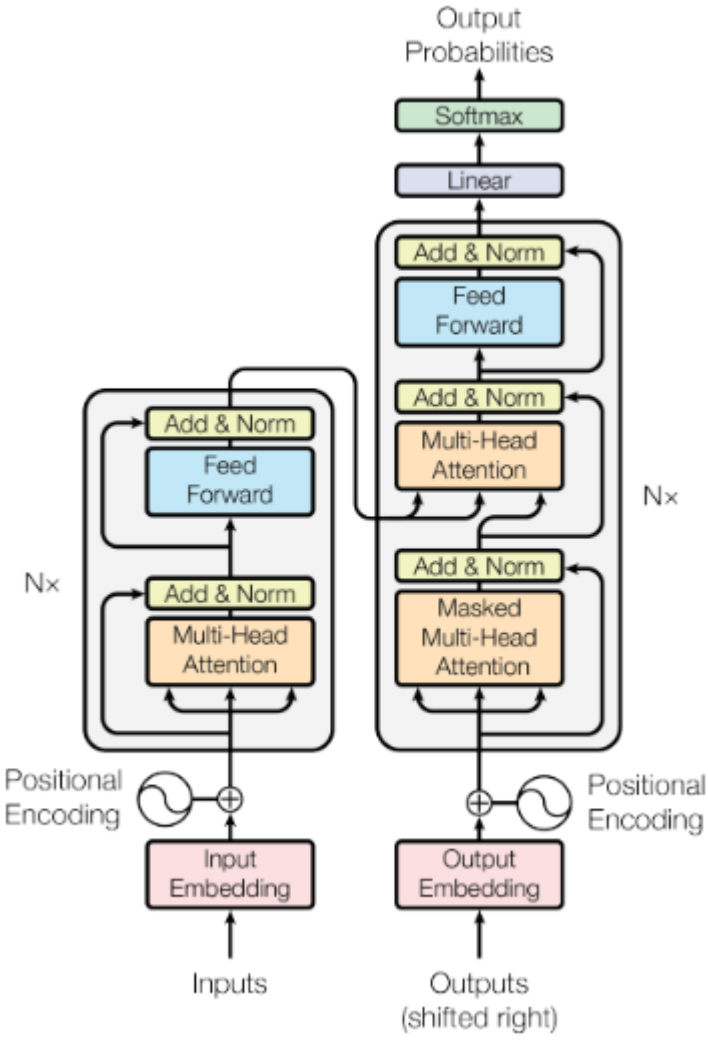


Рис. 7. Архитектура трансформера

После всех декодеров используется полносвязный слой Softmax, который преобразует полученные значения в вероятности, из которых затем выбирается наибольшее значение, а соответствующее ему слово становится выходом для этого временного шага.

АРХИТЕКТУРА GPT-2

GPT-2 – это большая языковая модель на основе модели Transformer, созданная некоммерческой компанией OpenAI, с количеством параметров от 117 миллионов до 1,5 миллиарда, обученная на наборе данных из 8 миллионов веб-страниц [20]. GPT-2 обучается с простой целью: предсказать следующее слово, учитывая все предыдущие слова в некотором тексте.

GPT-2 построена с использованием только блоков декодеров, которые имеют ту же структуру, что и в описанной выше общей модели Transformer.

В качестве входных данных GPT-2 использует не слова, а токены, полученные с помощью метода Byte Pair Encoding (BPE). Это метод сжатия данных, в котором наиболее распространенные пары последовательных байтов слов заменяются байтами, которых нет в этих словах [23]. Этот метод обеспечивает баланс между представлениями на уровне символов и слов, что позволяет ему обрабатывать большие корпуса данных.

Внутреннее внимание в GPT-2 также использует маскирование, которое блокирует информацию от токенов справа от вычисляемой позиции.

РЕАЛИЗАЦИЯ GPT-2

Была использована модель GPT-2 среднего размера с 345 млн параметров, состоящая из 24 блоков декодеров.

Модель была дообучена с помощью fine-tuning на корпусе поисковых запросов на английском языке, которые были также использованы для обучения рекуррентных нейронных сетей. С помощью полученной модели были сгенерированы поисковые запросы.

Была использована реализация модели, доступная по ссылке https://github.com/nshepperd/gpt-2. Модель была обучена на 1000 шагов.

ЛАТЕНТНО-СЕМАНТИЧЕСКИЙ АНАЛИЗ

Латентно-семантический анализ (Latent Semantic Analysis, LSA) – это метод обработки естественного языка для анализа зависимостей между коллекциями документов и терминами, содержащимися в них [24].

Метод использует терм-документную матрицу, которая описывает частоту появления терминов в коллекции документов. Элементы такой матрицы могут быть взвешены, например, с помощью TF-IDF: вес каждого элемента матрицы пропорционален количеству раз, когда термин встречается в каждом документе, и обратно пропорционален количеству раз, когда термин встречается во всех документах коллекции. После составления терм-документной матрицы проводится её сингулярное разложение, т. е. она представляется в виде $A = USV^T$, где мат-

рицы *U* и *V* – ортогональные, а *S* – диагональная матрица, значения которой называются сингулярными значениями матрицы *A*. Такое разложение отражает основную структуру зависимостей, присутствующих в исходной матрице, позволяя игнорировать шумы [25].

РЕАЛИЗАЦИЯ ЛАТЕНТНО-СЕМАНТИЧЕСКОГО АНАЛИЗА

Для проведения латентно-семантического анализа была использована библиотека gensim для Python [26]. Был создан корпус из 10000 документов, содержащий «эталонные» поисковые запросы, написанные людьми. Из него затем были удалены часто встречающиеся служебные слова английского языка (предлоги, артикли) и слова, встречающийся один раз, так как они не помогут вычислить семантическую связь между документами. С помощью класса Dictionary библиотеки gensim был создан словарь со словами и их индексами, затем с помощью метода doc2bow этого класса все документы представляются в формате «мешок слов» (bag of words). Модель TFIDF применена к полученному корпусу данных, а класс LsiModel проводит сингулярное разложение. Запросы, сгенерированные с помощью нейронных сетей, были токенизированы и с помощью словаря, созданного на «эталонном» корпусе, трансформированы в формат «мешок слов». Наконец, с помощью класса MatrixSimilarity вычисляются семантические сходства между этими корпусами с использованием косинусной меры.

РЕЗУЛЬТАТЫ ОЦЕНКИ СГЕНЕРИРОВАННЫХ ЗАПРОСОВ

Сравнивая каждый документ, в данном случае запрос, с документами из корпуса с реальными запросами, метод возвращает значение от –1 до 1, отражающее семантическое сходство документов. Результаты анализа приведены в таблице 2.

Таблица 2. Результаты латентно-семантического анализа

	GRU	LSTM	Fine-tuned GPT-2
Среднее значение	0,0065	0,006	0,0035
Среднее кол-во значений >0,7	16	14	9
Среднее кол-во значений > 0	4000	4659	2684

Корпус реальных запросов разнообразен, поэтому среднее значение результата сравнения каждого сгенерированного документа со всеми документами из «эталонного» корпуса незначительно отличается от нуля. При этом для каждого запроса, искусственно созданного с помощью сетей GRU и LSTM, существует в среднем 16 и 14 семантически близких документов, когда значения больше 0,7, а для дообученной модели GPT-2 это количество составило 9 документов. Также для каждого запроса, сгенерированного моделью GPT-2, из 10000 сравниваемых документов 2684 имеют значение больше 0, а для сетей LSTM и GRU – 4659 и 4000 соответственно. Из этого можно сделать вывод, что LSTM и GRU при генерации запросов использовали больше слов, семантически похожих на слова из обучающих данных, чем GPT-2. Это имеет смысл, так как первые две модели были обучены с нуля на входных данных, в то время как основное обучение последней модели происходило на совершенно ином корпусе, она была только дообучена с помощью метода fine-tuning, чтобы генерировать запросы, подходящие по структуре. Также важно принимать во внимание, что сравнение проводилось с 10 тысячами «эталонных» запросов, хотя модели обучались на 100 тысячах, соответственно не все зависимости были учтены, однако полученных значений достаточно для анализа.

Результаты анализа показывают, что сгенерированные запросы имеют схожую семантику с корпусом реальных запросов пользователей, но при этом не повторяют их буквально, то есть являются новыми по смыслу запросами.

Сети GRU и LSTM были обучены посимвольно и могли сгенерировать несуществующие слова, поэтому было решено проверить их. С помощью корпуса, содержащего более 466 тысяч английских слов, доступного по ссылке https://github.com/dwyl/english-words, каждое слово из запросов было проверено на существование. В запросах, сгенерированных сетью GRU, не было найдено 141 слово из 4431, а в запросах модели LSTM – 166 из 4325. Ненайденные слова содержали опечатки или ошибки, которые модели запомнили. Следовательно, возможно, стоит предобрабатывать данные, исправляя опечатки и ошибки такого рода. Однако запросы с опечатками могут быть полезны в зависимости от задачи, в которой они будут применяться. Так, например, при их использовании для те-

стирования новой поисковой системы или её оптимизации они будут более актуальными с опечатками, так как имеют большую схожесть с реальными пользовательскими запросами.

В силу того, что нейронные сети не могут понимать смысл предложения, хоть и часто находят верные зависимости между токенами, был проведен экспертный (ручной) анализ для оценки качества сгенерированных поисковых запросов.

Из запросов, созданных каждой моделью, случайным образом были выбраны 100 запросов. Было определено, имеет ли смысл каждый поисковый запрос и похож ли он на реальный возможный запрос пользователя. Следует отметить, что такая оценка субъективна. Запросы считались «хорошими», если слова в них были согласованы друг с другом.

Результаты анализа приведены в таблице 3.

Таблица 3. Результаты экспертной оценки поисковых запросов

	GRU	LSTM	Fine-tuned GPT-2
Хороший запрос	72	73	81
Плохой запрос	28	27	19

Из таблицы видно, что сети GRU и LSTM показали практически одинаковые результаты, а GPT-2 немного лучше. В ходе анализа было замечено, что модель GPT-2 генерирует более короткие запросы, чем две другие модели.

Результаты проведенных анализов показали, что сети GRU и LSTM имеют приблизительно одинаковое качество при решении задачи генерации поисковых запросов, а модель GPT-2 оказалась хуже в автоматическом анализе, но лучше при экспертной оценке. Следовательно, эта модель подходит лучше для генерации поисковых запросов, поскольку значимость экспертной оценки выше автоматической, хотя для более точных результатов стоит провести эту оценку с помощью других экспертов.

ЗАКЛЮЧЕНИЕ

В ходе работы исследованы ведущие модели, используемые для генерации текстов на естественном языке, для решения задачи генерации запросов к поисковым системам, а также проведен их сравнительный анализ. Полностью реализованы две нейронные сети: сеть с долгой кратковременной памятью и сеть с управляемым рекуррентным блоком. Исследована архитектура GPT-2, основанная на модели Transformer, она также была дообучена с помощью корпуса реальных запросов пользователей.

Латентно-семантический анализ показал, что модель GPT-2 имеет результаты хуже, чем две другие сети. Однако автоматические метрики оценки сгенерированного текста не всегда отражают качество модели, так как на данный момент невозможно оценить осмысленность текстов с помощью алгоритма. Для решения этой проблемы также был проведен экспертный анализ сгенерированных текстов, по результатам которого модель GPT-2 оказалась лучше двух других моделей. При этом сети LSTM и GRU показали приблизительно одинаковое качество по результатам всех проведенных анализов.

СПИСОК ЛИТЕРАТУРЫ

1. *Van Deemter K., Krahmer E., Theune M.* Real vs. template-based natural language generation: a false opposition?
URL: https://wwwhome.ewi.utwente.nl/~theune/PUBS/templates-squib.pdf

2. *Xie Z.* Neural Text Generation: A Practical Guide.
URL: https://arxiv.org/pdf/1711.09534.pdf

3. A Comprehensive Guide to Natural Language Generation, 2019.
URL: https://medium.com/sciforce/a-comprehensive-guide-to-natural-language-generation-dd63a4b6e548

4. *Arrington M.* AOL proudly releases massive amounts of user search data, 2006.
URL: https://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/

5. *Reiter E.* NLG vs Templates: Levels of Sophistication in Generating Text, 2016.
URL: https://ehudreiter.com/2016/12/18/nlg-vs-templates

6. *Gagniuc P.* Markov Chains: From Theory to Implementation and Experimentation, 2017. USA, NJ: John Wiley & Sons.

7. *Press O., Bar A., Bogin B., Berant J., Wolf L.* Language Generation with Recurrent Generative Adversarial Networks without Pre-training.
URL: https://arxiv.org/pdf/1706.01399.pdf

8. *Williams R.J., Hinton G.E., Rumelhart D.E.* Learning representations by back-propagating errors. URL: http://www.cs.utoronto.ca/~hinton/absps/naturebp.pdf

9. *Hochreiter S., Bengio Y., Frasconi P., Schmidhuber J.* Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies.
URL: https://www.bioinf.jku.at/publications/older/ch7.pdf

10. *Hochreiter S., Schmidhuber J.* Long-Short Term Memory.
URL: http://web.archive.org/web/20150526132154/;
URL: http://deeplearning.cs.cmu.edu/pdfs/Hochreiter97_lstm.pdf

11. *Heck J., Salem F.* Simplified Minimal Gated Unit Variations for Recurrent Neural Networks. URL: https://arxiv.org/abs /1701.03452

12. *Bahdanau D., Cho K.m Bengio Y.* Neural Machine Translation by Jointly Learning to Align and Translate. URL: https://arxiv.org/pdf/1409.0473.pdf

13. *Felbo B., Mislove A., Søgaard A., Rahwan I., Lehmann S.* Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. URL: https://arxiv.org/pdf/1708.00524.pdf

14. *Bisong E.* Google Colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform, 2019. Apress, Berkeley, CA.

15. *Chollet F. Keras,* 2015. URL: https://keras.io

16. *Kingma D., Ba J.* Adam: A Method for Stochastic Optimization.
URL: https://arxiv.org/abs/1412.6980

17. Learning Rate Scheduler.
URL: https://keras.io/api/callbacks/learning_rate_ scheduler/

18. *Schuster M., Paliwal K.* Bidirectional recurrent neural networks.
URL: https://www.researchgate.net/publication/3316656_Bidirectional_recurrent_neural_networks

19. *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaizer L., Polosukhin I.* Attention Is All You Need. URL: https://arxiv.org/pdf/1706.03762.pdf

20. *Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I.* Language Models Are Unsupervised Multitask Learners.
URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

21. *Devlin J., Chang M., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
URL: https://arxiv.org/pdf/1810.04805.pdf

22. *Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.* Language Models Are Few-Shot Learners. URL: https://arxiv.org/abs/2005.14165

23. *Gage P.* A New Algorithm for Data Compression.
URL: https://www.derczynski.com/papers/archive/BPE_Gage.pdf

24. *Deerwester S., Harshman R.* Indexing by Latent Semantic Analysis.
URL: https://www.cs.bham.ac.uk/ ~pxt/IDA/lsa_ind.pdf

25. *Nakov P.* Getting Better Results with Latent Semantic Indexing.
URL:http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.59.6406
&rep=rep1&type=pdf

26. *Rehurek R., Sojka P.* Software Framework for Topic Modelling with Large Corpora // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. University of Malta. 2010.

APPLYING MACHINE LEARNING TO THE TASK OF GENERATING SEARCH QUERIES

A. M. Gusenkov[1, [0000-0003-4019-7322]], A. R. Sittikova[2, [0000-0002-9539-764X]]

[1,2]Kazan Federal University

[1]gusenkov.a.m@gmail.com , [2]sitti.alina@mail.ru

Abstract

In this paper we research two modifications of recurrent neural networks – Long Short-Term Memory networks and networks with Gated Recurrent Unit with the addition of an attention mechanism to both networks, as well as the Transformer model in the task of generating queries to search engines. GPT-2 by OpenAI was used as the Transformer, which was trained on user queries. Latent-semantic analysis was carried

out to identify semantic similarities between the corpus of user queries and queries generated by neural networks. The corpus was converted into a bag of words format, the TFIDF model was applied to it, and a singular value decomposition was performed. Semantic similarity was calculated based on the cosine measure. Also, for a more complete evaluation of the applicability of the models to the task, an expert analysis was carried out to assess the coherence of words in artificially created queries.

Keywords*: natural language processing, natural language generation, machine learning, neural networks.*

REFERENCES

1. *Van Deemter K., Krahmer E., Theune M.* Real vs. template-based natural language generation: a false opposition?
URL: https://wwwhome.ewi.utwente.nl/~theune/PUBS/templates-squib.pdf

2. *Xie Z.* Neural Text Generation: A Practical Guide.
URL: https://arxiv.org/pdf/1711.09534.pdf

3. A Comprehensive Guide to Natural Language Generation, 2019.
URL: https://medium.com/sciforce/a-comprehensive-guide-to-natural-language-generation-dd63a4b6e548

4. *Arrington M.* AOL proudly releases massive amounts of user search data, 2006.
URL: https://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/

5. *Reiter E.* NLG vs Templates: Levels of Sophistication in Generating Text, 2016.
URL: https://ehudreiter.com/2016/12/18/nlg-vs-templates

6. *Gagniuc P.* Markov Chains: From Theory to Implementation and Experimentation, 2017. USA, NJ: John Wiley & Sons.

7. *Press O., Bar A., Bogin B., Berant J., Wolf L.* Language Generation with Recurrent Generative Adversarial Networks without Pre-training.
URL: https://arxiv.org/pdf/1706.01399.pdf

8. *Williams R.J., Hinton G.E., Rumelhart D.E.* Learning representations by back-propagating errors. URL: http://www.cs.utoronto.ca/~hinton/absps/naturebp.pdf

9. *Hochreiter S., Bengio Y., Frasconi P., Schmidhuber J.* Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies.

URL: https://www.bioinf.jku.at/publications/older/ch7.pdf

10. *Hochreiter S., Schmidhuber J.* Long-Short Term Memory.
URL: http://web.archive.org/web/20150526132154/;
URL: http:// deeplearning.cs.cmu.edu/pdfs/Hochreiter97_lstm.pdf

11. *Heck J., Salem F.* Simplified Minimal Gated Unit Variations for Recurrent Neural Networks. URL: https://arxiv.org/abs /1701.03452

12. *Bahdanau D., Cho K.m Bengio Y.* Neural Machine Translation by Jointly Learning to Align and Translate. URL: https://arxiv.org/pdf/1409.0473.pdf

13. *Felbo B., Mislove A., Søgaard A., Rahwan I., Lehmann S.* Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. URL: https://arxiv.org/pdf/1708.00524.pdf

14. *Bisong E.* Google Colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform, 2019. Apress, Berkeley, CA.

15. *Chollet F. Keras,* 2015. URL: https://keras.io

16. *Kingma D., Ba J.* Adam: A Method for Stochastic Optimization.
URL: https://arxiv.org/abs/1412.6980

17. Learning Rate Scheduler.
URL: https://keras.io/api/callbacks/learning_rate_ scheduler/

18. *Schuster M., Paliwal K.* Bidirectional recurrent neural networks.
URL: https://www.researchgate.net/publication/ 3316656_Bidirectional_recurrent_neural_networks

19. *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaizer L., Polosukhin I.* Attention Is All You Need. URL: https://arxiv.org/pdf/1706.03762.pdf

20. *Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I.* Language Models Are Unsupervised Multitask Learners.
URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

21. *Devlin J., Chang M., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
URL: https://arxiv.org/pdf/1810.04805.pdf

22. *Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.* Language Models Are Few-Shot Learners. URL: https://arxiv.org/abs/2005.14165

23. *Gage P.* A New Algorithm for Data Compression.
URL: https://www.derczynski.com/papers/archive/BPE_Gage.pdf

24. *Deerwester S., Harshman R.* Indexing by Latent Semantic Analysis.
URL: https://www.cs.bham.ac.uk/ ~pxt/IDA/lsa_ind.pdf

25. *Nakov P.* Getting Better Results with Latent Semantic Indexing.
URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.59.6406&rep=rep1&type=pdf

26. *Rehurek R., Sojka P.* Software Framework for Topic Modelling with Large Corpora // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. University of Malta. 2010.

СВЕДЕНИЯ ОБ АВТОРАХ

ГУСЕНКОВ Александр Михайлович – к. т. н., доцент Института вычислительной математики и информационных технологий Казанского (Приволжского) федерального университета. Области научных интересов: технологии извлечения знаний, обработка естественных языков, большие данные, интеллектуальный анализ данных.

Alexander M. GUSENKOV – assistant professor, Institute of Computational Mathematics and Information Technologies of Kazan Federal University. Ph.D. Current scientific interests: knowledge extraction technologies, Natural Language Processing, big data, data mining.

email: gusenkov.a.m@gmail.com

СИТТИКОВА Алина Рафисовна – бакалавр, Институт вычислительной математики и информационных технологий Казанского (Приволжского) федерального университета.

Alina R. SITTIKOVA – bachelor, Institute of Computational Mathematics and Information Technologies of Kazan Federal University.

sitti.alina@mail.ru

Материал поступил в редакцию 10 ноября 2020 года

УДК 004.822

ПРИНЦИПЫ ФОРМИРОВАНИЯ И ПРЕДСТАВЛЕНИЯ МЕЖДИСЦИПЛИНАРНЫХ КОЛЛЕКЦИЙ В ЦИФРОВОМ ПРОСТРАНСТВЕ НАУЧНЫХ ЗНАНИЙ

С. А. Кириллов[1, [0000-0001-7560-0041]], **И. Н. Соболевская**[2, [0000-0002-9461-3750]], **А. Н. Сотников**[3, [0000-0002-0137-1255]]

[1-3] *Межведомственный суперкомпьютерный центр Российской академии наук – филиал Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук» (МСЦ РАН – филиал ФГУ ФНЦ НИИСИ РАН), 119334 Москва, Ленинский проспект, 32а*

[1]skirillov@jscc.ru, [2]ins@jscc.ru, [3]ASotnikov@jscc.ru

Аннотация

Исследованы вопросы формирования междисциплинарных тематических коллекций в цифровом пространстве научных знаний. Рассмотрены содержание работ по формированию и представлению междисциплинарной коллекции, правила организации и представления междисциплинарных цифровых коллекций в информационной среде электронной библиотеки «Научное наследие России». Отмечено, что организация работ по формированию междисциплинарной коллекции в цифровом пространстве знаний предполагает следующие этапы: определение тематики междисциплинарной коллекции, определение структуры разделов междисциплинарной коллекции, определение источников для представления в междисциплинарной коллекции, диспетчеризацию работ с источниками, формирование метаданных, формирование цифровых копий объектов (включая оцифровку и верстку электронного объекта), размещение созданных цифровых копий на специализированной странице междисциплинарной коллекции. Показаны типы и виды междисциплинарных коллекций. Разработаны основные типы разделов, присутствующих в большинстве междисциплинарных проектов. Отмечено, что информация, представляемая в междисциплинарной коллекции, включает две составляющие – метаданные, описывающие характеристики ресурсов, и

собственно цифровые информационные ресурсы, а именно, представленные в цифровой форме объекты библиотечного, музейного и архивного хранения – это печатные и рукописные издания, графика, фото-, аудио-, видеоматериалы, музейные предметы. Предложена методика отбора материалов для формирования междисциплинарной коллекции на примере создания коллекции, посвященной нобелевским лауреатам, гражданам России и СССР, а также родившимся на территории России и СССР.

***Ключевые слова**: виртуальная выставка, электронная библиотека, научное наследие, базы данных, электронные фонды, цифровые копии.*

ВВЕДЕНИЕ

Под цифровым пространством научных знаний (ЦПНЗ) мы понимаем цифровую среду, при обращении к которой любой пользователь получит информацию касающиеся различных областей науки. Такая цифровая среда должна содержать достоверную информацию, основанную на фундаментальных научных знаниях [1].

Базис ЦПНЗ формируется из имеющихся библиотечных, архивных, музейных, энциклопедических, фактографических, словарных и других ресурсов. Этот контент создается научно-образовательным сообществом на основе существующих информационных систем [2, 3].

Одной из таких научных информационных систем является электронная библиотека «Научное наследие России». Междисциплинарные тематические проекты, реализованные средствами электронной библиотеки «Научное наследие России», позволяют интегрировать объекты различной природы (печатные издания, архивные документы, мультимедийные объекты), физически находящиеся в разных местах, в единый тематический ресурс и обеспечить его доступность пользователям.

1. МЕТОДИКА ФОРМИРОВАНИЯ И ПРЕДСТАВЛЕНИЯ МЕЖДИСЦИПЛИНАРНОЙ КОЛЛЕКЦИИ

Базовым архитектурным принципом формирования междисциплинарных коллекций выбран метод децентрализованной (распределенной) подготовки цифровых ресурсов и централизованного представления цифровых ресурсов на сайте или странице междисциплинарной коллекции. Применение распределенной схемы обусловлено тем, что распределены сами источники ресурсов (библиотеки, архивы, музеи), и сами объекты (книги, архивные материалы, фото-, видеодокументы, 3D модели предметов и т. п.) являются разнородными с точки зрения представления и описания метаданных.

Методика формирования междисциплинарной коллекции в цифровом пространстве знаний предполагает следующие этапы:

- определение темы междисциплинарной коллекции;
- определение структуры разделов междисциплинарной коллекции;
- определение и поиск источников для представления в междисциплинарной коллекции;
- формирование метаданных;
- формирование цифровых копий объектов, включая их оцифровку и верстку электронного объекта (создание электронной книги, 3D модели, оцифрованной копии фильма в требуемом цифровом формате);
- размещение созданных цифровых копий предметов библиотечного, архивного и музейного хранения, представленных в виде печатных изданий, рукописей, аудио-видео объектов, музейных предметов, других информационных материалов, на специализированной странице междисциплинарной коллекции.

Методика формирования междисциплинарных коллекций и их последовательность представлены на схеме 1.1.

Тема междисциплинарной коллекции должна быть актуальной и практически значимой для развития современной науки, соответствовать профилю цифрового пространства научных знаний. При утверждении темы междисциплинарной коллекции обосновываются актуальность темы, целевая аудитория потенциаль-

ных пользователей, выделяются базовые понятия, формируется перечень источников для формирования множества объектов формируемой коллекции.

Схема 1.1. Методика формирования междисциплинарной коллекции

По своему назначению междисциплинарные коллекции делятся на три типа:

- научно-исследовательские;
- научно-просветительные;
- образовательные, учебные.

Научно-исследовательские междисциплинарные коллекции объединяют результаты и осуществляют информационную поддержку фундаментальных и прикладных научных исследований.

Научно-просветительные междисциплинарные коллекции предназначены для распространения научных знаний.

Образовательные или учебные междисциплинарные коллекции создаются

для проведения различного рода занятий, связанных, например, с программами учебных заведений, и содержат: лекции, аудио и видео материалы и другой учебный контент [4].

Рассмотренные выше типы междисциплинарных коллекций состоят из следующих основных видов междисциплинарных коллекций:

- *персональная коллекция* посвящена научному наследию конкретного ученого;
- *тематическая коллекция* посвящена какому-либо научному направлению или научной проблеме;
- *событийная коллекция* посвящена особо важным событиям в истории российской науки (пример, Наука в СССР в годы Великой Отечественной войны 1941–1945 гг.);
- *корпоративная коллекция* освещает историю отечественных научных учреждений и обществ, научных школ;
- *справочная коллекция* содержит энциклопедическую и библиографическую информацию, архивные путеводители, описи и музейные каталоги.

Междисциплинарные коллекции, как правило, уникальны по структуре и содержанию. Вместе с тем работы по созданию междисциплинарных коллекций позволили разработать основные типы разделов, присутствующих в большинстве проектов, а именно:

- основной тематический раздел;
- биографический раздел;
- интерактивный раздел;
- раздел видеоматериалов;
- раздел фотодокументов;
- библиотека;
- раздел «Коллекция 3D объектов»;
- раздел отзывов;
- контакты.

Основной тематический раздел. Здесь содержится подборка материалов, раскрывающих основную тему междисциплинарной коллекции. Как правило, это

статьи (в том числе авторские, написанные специально для раскрытия тематики коллекции), подборки редких фотоматериалов и изображений, выдержки из публицистических и периодических изданий.

Биографический раздел. Этот раздел содержит биографические справки и портреты ученых.

Интерактивный раздел. Управляющие элементы интерфейса этого раздела должны дать возможность пользователю взаимодействовать с элементами виртуальной экспозиции. Задача раздела – повышение заинтересованности пользователя (посетителя виртуальной выставки), переход от пассивного восприятия информации к активному пониманию коллекции. Основными элементами интерактивного раздела являются научные викторины, интеллектуальные игры, 3D анимации.

Раздел видеоматериалов. Видео материалы в этом разделе представлены документальными фильмами, архивными видеоматериалами и/или научно-популярными фильмами. Просмотр видеоматериалов реализован как в режиме предпросмотра, так и в полноэкранном режиме. Также реализуются все необходимые элементы управления для просмотра видео. Кроме материалов, предоставленных участниками проекта, в разделе кинодокументов формируются ссылки на видеоматериалы, находящиеся в свободном доступе в интернете, если таковые имеются.

Раздел фотодокументов. Раздел, как правило, содержит уникальные фотодокументы, предоставленные участниками проекта.

Электронная библиотека. В разделе представлены публикации по заданной тематике из фондов электронной библиотеки «Научное наследие России». Раздел оформляется в виде интерактивного списка авторов и публикаций, пользуясь которым читатель попадает непосредственно на страницу, посвященную ученому, или в его публикацию на сайте электронной библиотеки «Научное наследие России». Дополнительно представлены ссылки на издания, находящиеся в свободном доступе в интернете. Для изданий, которые по каким-либо причинам еще не оцифрованы, создаются библиографические списки в формате, позволяющем пользователям коллекции найти и прочитать эти книги в других научных библиотеках.

Раздел «Коллекция 3D». Раздел представляет собой галерею, состоящую из 3D макетов оцифрованных архивных или музейных предметов. Предметы оцифрованы таким образом, чтобы пользователь междисциплинарной коллекции смог детально рассмотреть предмет во всех ракурсах.

Раздел отзывов. Данный раздел создается для обратной связи с посетителями сайта. Здесь можно (после обязательной регистрации) обменяться мнениями, выступить с различными сообщениями по тематике выставки.

Раздел «контакты». Здесь указываются контакты администратора виртуальной выставки для связи с ним.

Источниками объектов для представления в междисциплинарной коллекции являются фонды предметов библиотечного, архивного, музейного хранения, представленных в виде цифровых копий печатных изданий, рукописей, аудио-видео объектов, 3D макетов оцифрованных архивных или музейных предметов, биографических справок и других информационных ресурсов. Отобранные объекты объединяются определенным общим набором свойств и обладают определенной тематической связанностью [5].

Для подбора источников, представленных в междисциплинарной коллекции, необходимо выполнять положения законов о защите авторских прав. На издания, защищенные законодательством, требуется получить письменные согласия владельцев авторских прав.

Система диспетчеризации обеспечивает сопровождение и контроль процесса выполнения технологического цикла отбора и создания элементов междисциплинарной коллекции. С помощью набора управляющих инструкций в системе диспетчеризации на каждом этапе фиксируется текущее состояние работ, что обеспечивает контроль их выполнения.

В распределенной системе это позволяет получать оперативную информацию о ходе работ по формированию коллекции (как по ученым, так и по публикациям), справки о текущем состоянии процессов обработки конкретного издания, числе сверстанных изданий и страниц, количественные данные о работе, выполненной участниками проекта.

Информация, представляемая в междисциплинарной коллекции, включает две составляющие – метаданные, описывающие характеристики ресурсов, и соб-

ственно цифровые информационные ресурсы, а именно, представленные в цифровой форме объекты библиотечного, музейного и архивного хранения (это печатные и рукописные издания, графика, фото-, аудио-, видеоматериалы, музейные предметы).

Метаданные – это структурированные данные, которые описывают характеристики объектов-носителей информации, способствующие идентификации, обнаружению, оценке и управлению этими объектами. Также метаданные представляют собой совокупность формальных признаков, по которым осуществляются описание и поиск цифровых ресурсов (пример: для публикации – автор, заглавие, ключевые слова, год издания, место издания и т. п.).

Формирование цифровых копий объектов выполняется на различном специализированном оборудовании. Для перевода книг, карт, рукописей, различных графических изображений в цифровую форму используются планетарные, планшетные или иные типы сканеров, различающиеся между собой разрешающей способностью, возможностью сканирования различного типа изображений и другими свойствами. Для оцифровки объемных объектов применяются 3D сканеры либо специализированное фотооборудование. Файлы, полученные в результате оцифровки, передаются в группу верстки цифрового ресурса. На этом этапе из набора отдельных изображений, после соответствующей обработки, осуществляется создание (верстка) цифрового ресурса. В результате верстки мы получаем электронную книгу, электронный 3D образ предмета, оцифрованную копию фильма или фотографии в требуемом формате [6].

На заключительном этапе, после проверки, оцифрованная информация и метаданные поступают на сайт междисциплинарной коллекции.

2. ОТБОР МАТЕРИАЛОВ ДЛЯ ФОРМИРОВАНИЯ МЕЖДИСЦИПЛИНАРНОЙ КОЛЛЕКЦИИ

Рассмотрим методику отбора материалов для формирования междисциплинарной коллекции на примере создания коллекции, посвященной нобелевским лауреатам, гражданам России и СССР, а также родившимся на территории России и СССР.

На первом этапе необходимо сформировать список персон, удовлетворяющих условиям: «Человек является или являлся гражданином России и нобелевским лауреатом»; «Человек являлся гражданином СССР на момент получения премии и нобелевским лауреатом»; «Человек является подданным России (Российской империи) на момент получения премии и нобелевским лауреатом»; «Человек родился на территории России (Российской империи) или СССР, является нобелевским лауреатом». Такой список можно получить, например, из [7]. В таблице 1 приведен список нобелевских лауреатов, подданных России и СССР на момент получения премии:

Таблица 1. Список нобелевских лауреатов, подданных России и СССР на момент получения премии

№	Год	Направление	Лауреат
1	1904	Физиология и медицина	Павлов Иван Петрович
2	1905	Литература	Сенкевич Генрик
3	1908	Физиология и медицина	Мечников Илья Ильич
4	1933	Литература	Бунин Иван Алексеевич
5	1956	Химия	Семёнов Николай Николаевич
6	1958	Литература	Пастернак Борис Леонидович
7	1958	Физика	Черенков Павел Алексеевич
8	1958	Физика	Тамм Игорь Евгеньевич
9	1958	Физика	Франк Илья Михайлович
10	1962	Физика	Ландау Лев Давидович
11	1964	Физика	Басов Николай Геннадьевич
12	1964	Физика	Прохоров Александр Михайлович
13	1965	Литература	Шолохов Михаил Александрович
14	1970	Литература	Солженицын Александр Исаевич
15	1975	Экономика	Канторович Леонид Витальевич
16	1975	Премия мира	Сахаров Андрей Дмитриевич
17	1978	Физика	Капица Пётр Леонидович
18	1990	Премия мира	Горбачёв Михаил Сергеевич
19	2000	Физика	Алфёров Жорес Иванович

20	2003	Физика	Абрикосов Алексей Алексеевич
21	2003	Физика	Гинзбург Виталий Лазаревич
22	2010	Физика	Новосёлов Константин Сергеевич

Таблица 2. Список нобелевских лауреатов, родившихся на территории России и СССР

№	Год	Направление	Лауреат	Место рождения	Гражданство на момент получения премии
1	1903	Физика	Склодовская-Кюри Мария	Варшава	Франция
2	1909	Химия	Оствальд Вильгельм	Рига	Германия
3	1911	Химия	Склодовская-Кюри Мария	Варшава	Франция
4	1924	Литература	Реймонт Владислав	Кобелех-Вельких	Польша
5	1937	Химия	Каррер Пауль	Москва	Швейцария
6	1939	Литература	Силланпяя Франс Эмиль	Хямеэнкурё	Финляндия
7	1945	Химия	Виртанен Арттури Илмари	Гельсингфорс	Финляндия
8	1950	Химия	Рейхштейн Тадеуш	Влоцлавк	Швейцария
9	1952	Физиология и медицина	Ваксман Зельман	Новая Прилука	США
10	1967	Физиология и медицина	Гранит Рагнар	Рийхимяки	Швеция
11	1971	Экономика	Кузнец Саймон	Пинск	США

12	1973	Экономика	Леонтьев Василий	Мюнхен (по рождению подданный Российской империи)	США
13	1977	Химия	Пригожин Илья	Москва	Бельгия
14	1978	Литература	Зингер Исаак Башевис	Леончин	США
15	1978	Премия мира	Бегин Менахем	Брест-Литовск	Израиль
16	1980	Литература	Милош Чеслав	Шетенях	Польша
17	1987	Литература	Бродский Иосиф Александрович	Ленинград	США
18	1995	Премия мира	Ротблат Джозеф	Варшава	Великобритания
19	2007	Экономика	Гурвич Леонид	Москва	США
20	2010	Физика	Гейм Андрей Константинович	Сочи	Нидерланды
21	2015	Литература	Алексиевич Светлана Александровна	Станислав	Белоруссия

На втором этапе необходимо определить наличие или отсутствие в электронной библиотеке, на базе которой будет формироваться коллекция, сведений о каждой из персон, приведенных в таблицах 1 и 2, а также наличие публикаций либо самих персон, либо о них.

В данной работе приведен пример формирования междисциплинарной коллекции на платформе электронной библиотеки «Научное наследие России» (ЭБ ННР). В таблице 3 представлены вышеуказанные сведения:

Таблица 3. Количество работ и сведений о персонах, приведенных в таблицах 1 и 2, в ЭБ «Научное наследие России».

№	Год	Направление	Лауреат	Число работ ученого	Работы об ученом
1	1904	Физиология и медицина	Павлов Иван Петрович	12	10
2	1905	Литература	Сенкевич Генрик	нет	
3	1908	Физиология и медицина	Мечников Илья Ильич	31	4
4	1933	Литература	Бунин Иван Алексеевич	нет	
5	1956	Химия	Семёнов Николай Николаевич	14	
6	1958	Литература	Пастернак Борис Леонидович	нет	
7	1958	Физика	Черенков Павел Алексеевич	1	
8	1958	Физика	Тамм Игорь Евгеньевич	5	
9	1958	Физика	Франк Илья Михайлович	1	
10	1962	Физика	Ландау Лев Давидович	15	
11	1964	Физика	Басов Николай Геннадьевич		
12	1964	Физика	Прохоров Александр Михайлович		
13	1965	Литература	Шолохов Михаил Александрович	нет	

14	1970	Литература	Солженицын Александр Исаевич	нет	
15	1975	Экономика	Канторович Леонид Витальевич	8	
16	1975	Премия мира	Сахаров Андрей Дмитриевич		
17	1978	Физика	Капица Пётр Леонидович		1
18	1990	Премия мира	Горбачёв Михаил Сергеевич	нет	
19	2000	Физика	Алфёров Жорес Иванович	нет	
20	2003	Физика	Алексей Алексеевич Абрикосов	нет	
21	2003	Физика	Виталий Лазаревич Гинзбург	3	
22	2010	Физика	Константин Сергеевич Новосёлов	нет	
ИТОГО					

Прежде, чем перейти к третьему этапу, необходимо отметить, что в данной работе рассматривается пример формирования научной коллекции, поэтому из приведенных выше списков нобелевских лауреатов исключим тех, кто получил премию по литературе и премию мира, не будучи ученым в области естественных наук.

На третьем этапе необходимо отобрать кино- и фотодокументы, хранящиеся в Российском государственном архиве кинофотодокументов (РГАКФД), связанные с деятельностью ученых. В таблице 4 приведен список тех ученых, с которыми связана кино-, видео- или фотохроника в РГАКФД. Эта информация может быть получена с помощью электронного каталога на официальном сайте РГАКФД [8]. В таблице 4 приведена информация о наличии кино-, видео- или фотодокументов, так или иначе связанных с приведенными выше персонами.

Таблица 4. Информация о наличии в РГАКФД кино-, видео- или фотодокументов, так или иначе связанных с приведенными выше персонами.

№	Лауреат	Кинохроника	Фотоматериалы	Видеохроника
1	Павлов Иван Петрович	41	1	1
2	Мечников Илья Ильич	5	1	
3	Семёнов Николай Николаевич	16	17	
4	Черенков Павел Алексеевич	10	2	
5	Тамм Игорь Евгеньевич	24	2	2
6	Франк Илья Михайлович	18	1	
7	Ландау Лев Давидович	7	1	7
8	Басов Николай Геннадьевич	17	3	3
9	Прохоров Александр Михайлович	15		2
10	Канторович Леонид Витальевич	1	2	
11	Сахаров Андрей Дмитриевич	38	8	21
12	Капица Пётр Леонидович	45	5	14
13	Алфёров Жорес Иванович			13
14	Алексей Алексеевич Абрикосов	1		
15	Виталий Лазаревич Гинзбург	2		1
16	Константин Сергеевич Новосёлов			
17	Склодовская-Кюри Мария	1		1

На четвертом этапе необходимо определить, где могут храниться документы, архивы или музейные предметы, связанные с данными персонами. Такие материалы могут находиться, в том числе, в научных учреждениях, с которыми была связана деятельность ученого, музеях-квартирах, личных архивах наследников и т. п.

На пятом этапе необходимо связаться с владельцами данных ресурсов и заключить соглашение о сотрудничестве, на основании которого будет сформирована совместная выставка, посвященная нобелевским лауреатам, родившимся на территории России или СССР.

ЗАКЛЮЧЕНИЕ

Формирование цифрового пространства научных знаний является одним из важнейших объектов современного информационного общества. Такое пространство содержит элементы, представляющие собой некоторые формализованные характеристики научного знания и образует множество подпространств информационных систем, создаваемых для предоставления доступа к предметной области или профессиональным знаниям, при обращении к которым любой пользователь, будь то учёный или ученик средней школы, получит ответы на вопросы, касающиеся различных областей науки. Частным случаем такого цифрового пространства может выступать, в том числе, информационная среда электронной библиотеки (ЭБ) [9].

Пространство такой электронной библиотеки должно формироваться на основе децентрализованной подготовки метаданных. Эти метаданные должны создаваться по единым правилам с централизованным хранением при единой централизованной системе контроля качества. Оцифрованные публикации, архивные документы, музейные предметы и т. п. могут храниться как у владельцев этих ресурсов, так и в центральном блоке пространства ЭБ.

Такая организация ЭБ позволит решить, в том числе, задачи интеграции и представления информационных объектов различной природы (печатные издания, архивные документы, мультимедийные объекты).

Описанный выше алгоритм формирования междисциплинарной коллекции позволяет хранить и предоставлять пользователю эту коллекцию как «саму по себе», так и формировать на ее основе междисциплинарные выставки [10].

Описанная выше технология легла в основу нескольких междисциплинарных проектов. Одним из этих проектов является виртуальная выставка, посвященная 160-летию со дня рождения И.В. Мичурина. Эта выставка создана совместно с Государственным биологическим музеем имени К.А. Тимирязева (ГБМТ) и Российским Государственным архивом кинофотодокументов (РГАКФД)

(http://vim.benran.ru/). Выставка посвящена не только биографии и научной деятельности И.В. Мичурина, но и истории развития генетики в СССР в целом. В рамках работы над проектом создано и представлено широкому кругу пользователей более 70 макетов плодов И.В. Мичурина, хранящихся в запасниках ГБМТ, также на выставке представлены уникальные кадры фото- и кинохроники, связанные с жизнедеятельностью И.В. Мичурина.

Другая выставка, также сформированная по описанной выше технологии, посвящена М.М. Герасимову и его антропологическим реконструкциям (http://acadlib.ru/). Этот проект создан совместно с ГБМТ, РГАКФД и Государственным Дарвиновским музеем. В рамках работы над проектом создано и представлено широкому кругу пользователей более 50 антропологических реконструкций М.М. Герасимова.

Обе выставки реализованы на «самостоятельных» платформах, а также интегрированы в ЭБ «Научное наследие России».

В настоящее время выставочные проекты активно развиваются, дополняются новыми материалами. Кроме того, планируется создание виртуальных аудио экскурсий по ним.

Благодарности

Работа выполнена при поддержке РФФИ, проект № 20-07-00773.

СПИСОК ЛИТЕРАТУРЫ

1. *Антопольский А.Б., Каленов Н.Е., Серебряков В.А., Сотников А.Н.* О едином цифровом пространстве научных знаний // Вестник Российской академии наук. 2019. Т. 89, № 7. С. 728–735. http://dx.doi.org/10.31857/S0869-5873897728-735.

2. *Antopolskii A.B.* Future of Scientific Communications and Scientific Information // Information and Innovation. 2019. V. 14, No. 1. P. 7–17.

3. *Zhmailo S.V., Ulyanin O.V.* Sci-tech libraries within the knowledge management system: from information specialist's viewpoint // Nauchnye i tekhnicheskie biblioteki (Scientific and Technical Libraries). 2020. No. 2. P. 9–23.

4. *Каленов Н.Е., Соболевская И.Н., Сотников А.Н.* Hierarchical representation of information objects in a digital library environment // 17th Russian Conference,

RCAI 2019, Ulyanovsk, Russia, October 21–25, 2019, Proceedings, ISSN 1865-0929, P. 93–104. URL: https://link.springer.com/chapter/10.1007/978-3-030-30763-9_8.

5. *Sobolevskaya I.N., Sotnikov A.N.* Principles of 3D Web-collections Visualization // Proceedings of the 3rd International Conference on Computer-Human Interaction Research and Application. ISSN 978-989-758-376-6. 2019. P. 145–151.

6. *Cooper J.P., Wetherelt A., Zazzaro Ch*. From Boatyard to Museum: 3D laser scanning and digital modelling of the Qatar Museums watercraft collection //International Journal of Nautical Archaeology». 2018. V. 47, No. 2. P. 419–442.

7. Нобелевские лауреаты из России и СССР.
URL: https://ru.wikipedia.org/wiki/ Нобелевские_лауреаты_из_России_и_СССР

8. Российский государственный архив кинодокументов. URL: http://rgakfd.ru/

9. *Власова С.А., Каленов Н.Е.* Интернет-каталог Библиотеки по естественным наукам Российской академии наук как специальная информационно-поисковая система, ориентированная на квалифицированного пользователя // Системы и средства информатики. 2019. Т. 29, № 1. С. 86–95.
http://dx.doi.org/10.14357/08696527190108.

10. *Сотников А.Н., Соболевская И.Н.* An example of the formation of a digital exhibition space with the means of the virtual exhibition "anthropological reconstructions. M.M. Gerasimov's scientific heritage" // Information Innovative Technologies: Materials of the International scientific-practical conference /Ed. Uvaysov S.U., Ivanov I.A. M.: Association of graduates and employees of AFEA named after prof. Zhukovsky, 2019. ISSN 2542-1824. P. 12–17.
URL: https://cloud.mail.ru/public/3CVd/5L3MjJzj5

SOME ASPECTS OF THE FORMATION AND REPRESENTATION PRNCIPLE OF INTERDISCIPLINARY COLLECTION IN THE DIGITAL SPACE OF SCIENTIFIC KNOWLEDGE

S. A. Kirillov[1, [0000-0001-7560-0041]] , I. N. Sobolevskaya[2, [0000-0002-9461-3750]], A. N. Sotnikov[3, [0000-0002-0137-1255]]

[1 – 3] *Joint Super Computer Center of the Russian Academy of Sciences – Branch of Federal State Institution "Scientific Research Institute for System Analysis of the Russian Academy of Sciences", 119334 Moscow, Leninsky Prospect, 32a*

[1]skirillov@jscc.ru, [2]ins@jscc.ru, [3] ASotnikov@jscc.ru

Abstract

Interdisciplinary thematic projects implemented by means of the electronic library "Scientific heritage of Russia" allow integrating objects of various nature (printed publications, archival documents, multimedia objects) into a single thematic resource and making it accessible to users. The approaches to the formation of interdisciplinary thematic collections in the digital space of scientific knowledge are investigated. Algorithms for the formation and presentation of a digital interdisciplinary collection are presented. The method of creation and presentation of virtual collections in the information environment of the electronic library "Scientific heritage of Russia". The main types of sections present in most projects are indicated. The main stages of the formation of an interdisciplinary collection in the digital space of knowledge have been formed and described, including the composition of the collection sections, sources for presenting collection materials, dispatching work with sources, the formation of metadata, the main types of sections, etc. An example of the application of the content formation methodology for creating an interdisciplinary collection is given.

Keywords*: virtual exhibition, e-library, scientific heritage, databases, electronic records, digital copies.*

REFERENCES

1. *Antopol'skij A.B., Kalenov N.E., Serebryakov V.A., Sotnikov A.N.* O edinom cifrovom prostranstve nauchnyh znanij // Vestnik Rossijskoj akademii nauk, 2019. T. 89, № 7. S. 728–735. http://dx.doi.org/10.31857/S0869-5873897728-735

2. *Antopolskii A.B.* Future of Scientific Communications and Scientific Information. Information and innovation. 2019. V. 14. No. 1. S. 7–17.

3. *Zhmailo S.V., Ulyanin O.V.* Sci-tech libraries within the knowledge management system: from information specialist's viewpoint. Nauchnye i tekhnicheskie biblioteki (Scientific and Technical Libraries). 2020. V. 2. S. 9–23.

4. *Kalenov N.E., Sobolevskaya I.N., Sotnikov A.N.* Hierarchical representation of information objects in a digital library environment // 17th Russian Conference, RCAI 2019, Ulyanovsk, Russia, October 21–25, 2019, Proceedings, ISSN 1865-0929, P. 93–104, URL: https://link.springer.com/chapter/10.1007/978-3-030-30763-9_8.

5. *Sobolevskaya I.N., Sotnikov A.N.* Principles of 3D Web-collections Visualization // Proceedings of the 3rd International Conference on Computer-Human Interaction Research and Application. ISSN: 978-989-758-376-6. 2019. P. 145–151.

6. *Cooper J.P., Wetherel A., Zazzaro Ch*. From Boatyard to Museum: 3D laser scanning and digital modelling of the Qatar Museums watercraft collection. «International journal of nautical archaeology». 2018. V. 47. No. 2. P. 419–442.

7. List of Russian Nobel laureates. URL: https://en.wikipedia.org/wiki/List_of_Russian_Nobel_laureates

8. Russian State Documentary Film & Photo Archive. URL: http://rgakfd.ru/

9. *Vlasova S.A., Kalenov N.E.* Internet-katalog Biblioteki po estestvennym naukam Rossijskoj akademii nauk kak special'naya informacionno-poiskovaya sistema, orientirovannaya na kvalificirovannogo pol'zovatelya // Sistemy i sredstva informatiki, 2019. V. 29. № 1. S. 86–95. http://dx.doi.org/10.14357/08696527190108.

10. *Sotnikov A.N., Sobolevskaya I.N.* An example of the formation of a digital exhibition space with the means of the virtual exhibition "anthropological reconstructions. M.M. Gerasimov's scientific heritage", Information Innovative Technologies: Materials of the International scientific – practical conference / Uvaysov S.U., Ivanov I.A. (Eds.) M.: Association of graduates and employees of AFEA named after prof. Zhukovsky. 2019. P. 12–17. URL: https://cloud.mail.ru/public/3CVd/5L3MjJzj5

СВЕДЕНИЯ ОБ АВТОРАХ

КИРИЛЛОВ Сергей Александрович – старший научный сотрудник Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук». Сфера научных интересов – программные средства и системы для распределенных вычислений; формирование баз данных для электронных библиотек; методы, средства и системы обработки данных большого объема; 3D-моделирование.

Sergey Aleksandrovich KIRILLOV – senior scientist researcher of Joint SuperComputer Center of the Russian Academy of Sciences – Branch of Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences”. Research interests include software and systems for distributed computing; e-library database building; methods, tools and systems of large data processing; 3D modeling.

email: skirillov@jscc.ru

СОБОЛЕВСКАЯ Ирина Николаевна– старший научный сотрудник Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», к. ф.-м. н. Сфера научных интересов – математическое обеспечение, программные средства и системы для распределенных вычислений; формирование баз данных для электронных библиотек; методы, средства и системы обработки данных большого объема; 3D-моделирование.

Irina Nikolaevna SOBOLEVSKAYA – senior scientist researcher of Joint SuperComputer Center of the Russian Academy of Sciences – Branch of Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences”. Research interests include mathematical software, software and systems for distributed computing; e-library database building; methods, tools and systems of large data processing; 3D modeling.

email: ins@jscc.ru

СОТНИКОВ Александр Николаевич – зам. директора по научной работе Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», д. ф.-м. н., профессор. Сфера научных интересов – математическое обеспечение, программные средства и системы для распределенных вычислений; формирование баз данных для электронных библиотек; методы, средства и системы обработки данных большого объема; нейронные и семантические сети.

Aleksandr Nikolaevich SOTNIKOV – Deputy director for science of Joint SuperComputer Center of the Russian Academy of Sciences – Branch of Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences”. Research interests include mathematical software, software and systems for distributed computing; e-library database building; methods, tools and systems of large data processing; semantic and nerve nets.

email: ASotnikov@jscc.ru

Материал поступил в редакцию 27 ноября 2020 года

УДК 004.912

ИСПОЛЬЗОВАНИЕ МЕТОДОВ ТЕМАТИЧЕСКОГО АНАЛИЗА В НАУКОМЕТРИЧЕСКИХ СИСТЕМАХ

А. С. Козицын[1, [0000-0002-8065-9061]], С. А. Афонин[2, [0000-0003-3058-9269]], Д. А. Шачнев[3, [0000-0002-5940-9180]]

НИИ механики МГУ им. М.В. Ломоносова

[1]alexanderkz@mail.ru, [2]serg@msu.ru, [3]mitya57@gmail.com

Аннотация

Во многих современных наукометрических системах и системах цитирования представлены различные механизмы тематического поиска и тематической фильтрации информации. В большинстве случаев для тематического анализа статей и журналов используется полнотекстовый подход, который имеет ряд ограничений. Использование алгоритмов, основанных на анализе графов как автономно, так и совместно с полнотекстовыми алгоритмами, позволяет устранить эти ограничения и улучшить полноту и точность тематического поиска. Алгоритм, разработанный авторами и представленный в этой работе, использует для анализа тематической близости журналов граф соавторства. Алгоритм нечувствителен к языку журнала и подбирает похожие журналы на разных языках, что сложно реализуемо для алгоритмов, основанных на анализе полнотекстовой информации. Апробация алгоритма проводилась в наукометрической системе ИАС ИСТИНА. В интерфейсе, разработанном для этих целей, пользователь может выбрать один близкий ему по тематике журнал, и система автоматически сформирует подборку журналов, которые могут представлять интерес для пользователя как с точки зрения изучения имеющихся в них материалов, так и с точки зрения публикации собственных статей. В перспективе разработанный алгоритм можно адаптировать для поиска похожих по тематике конференций, сборников публикаций и научных проектов. Наличие такого инструмента увеличит публикационную активность молодых сотрудников, повысит цитируемость статей и цитируемость между журналами. Результаты работы алгоритма определения тематической близости между

журналами, сборниками, конференциями и научными проектами также могут использоваться для построения правил в моделях разграничения доступа к данным на основе онтологий предметной области.

Ключевые слова*: тематическая классификация, библиографические данные, граф соавторства, информационные системы*.

ВВЕДЕНИЕ

Применение современных методов тематического анализа для аналитической обработки больших объемов информации используется в настоящее время практически во всех сферах человеческой деятельности, в том числе, в наукометрии. Результаты тематического анализа научной информации могут использоваться в целях уточнения наукометрических показателей, принятия управленческих решений, информационного поиска и определения правил доступа к информации.

Расчет наукометрических показателей используется для оценки значимости статей (цитируемость), авторитетности журналов (импакт-фактор, h5-индекс, h5-медиана), влияния на научное сообщество отдельных авторов (индекс Хирша и g-индекс), оценки деятельности организаций в целом (i-индекс) [1]. Однако многими авторами отмечается, что характеристики распределения абсолютных числовых значений наукометрических показателей имеют существенную зависимость от анализируемой тематической области [2]. Например, значения индекса цитируемости статей за последние 2 года имеют различную медиану для физики и математики, поскольку математические статьи дольше цитируются, но медленнее «набирают» количество ссылок. Аналогичное несоответствие показателей наблюдается и в журналах в целом. Например, лучшие российские журналы по данным РИНЦ, представленные на странице статистики elibrary.ru/titles_compare.asp, за 2018-й год по состоянию на 20.04.2020 имеют следующие показатели цитируемости по разным рубрикам: физика 9200; биология 4600; математика 3600; механика 1500; информатика 1100. В этой связи проводить сравнение по абсолютным значениям наукометрических показателей статей, журналов или авторов из разных тематических направлений некорректно. Необходимо в таких случаях использовать нормализованную среднюю цитируе-

мость [3] или другие аналогичные показатели, учитывающие тематическую область проводимых исследований. Построение таких нормализованных показателей требует проведения тематической классификации больших объемов научных статей и журналов.

При осуществлении управленческой деятельности использование результатов тематического анализа позволяет оценивать состояние различных направлений исследований, проводить их сравнение с мировым уровнем, выявлять новые тематические направления для определения политики выделения материальных ресурсов для стимулирования научной деятельности. При этом необходимо оценивать не только текущие значения показателей, но и их динамику во времени, а также мировые показатели. Например, уменьшение показателей по определенной тематике исследований на фоне роста этих же показателей в мире может означать отток научных кадров в организации из этой области исследований или устаревание оборудования.

Еще одним важным направлением применения тематического анализа является создание эффективных механизмов для проведения информационного поиска. Объектами поиска могут являться: публикации, журналы, персоны, организации и другие объекты. На основе проведения тематической классификации и кластеризации могут решаться такие актуальные задачи, как поиск опубликованных материалов по заданной тематике, поиск наиболее авторитетных экспертов в определенной предметной области, определение списка журналов для публикации и оценка их значимости, выделение новых тематических направлений в какой-либо области и поиск научных коллективов.

Определение тематических связей между объектами информационной системы [4] также может использоваться для автоматического построения онтологий и определения правил доступа к данным в моделях логического разграничения доступа ABAC (Attribute-Based Access Control) [5], которые в настоящее время в значительной степени потеснили старые модели разграничения доступа: ролевую модель RBAC; мандатную модель MAC и дискреционную модель DAC.

Многие крупные наукометрические системы и системы цитирования имеют инструментарий, позволяющий проводить тематический анализ данных.

1. ИСПОЛЬЗОВАНИЕ ТЕМАТИЧЕСКОГО АНАЛИЗА В СОВРЕМЕННЫХ СИСТЕМАХ

Возможности тематического анализа различных наукометрических систем различаются по типу обрабатываемой информации, видам классификаторов, источникам информации, набору используемых методов классификации и кластеризации.

Проект Web of Science (WoS) для проведения тематического поиска использует индексы по ключевым словам и тематические классификаторы. Индексация по ключевым словам производится с использованием авторских ключевых слов (Author Keywords), которые авторы указывают вручную при добавлении статьи. Также производится индексация по ключевым словам и терминам (KeyWords Plus), автоматически выделенным из названий статей, цитируемых в работе. Индексация по ключевым словам позволяет производить поиск и дополнительную фильтрацию с использованием терминов, заданных пользователем. Для индексации по тематическим классификаторам используется два основных классификатора: одноуровневый классификатор Web of Science Categories для журналов, содержащий 250 категорий, и двухуровневый классификатор статей Research Area по 150 областям науки. Кроме того, используется дополнительный классификатор Essential Science Indicators из 22 категорий. В проекте реализован сервис Manuscript Matcher, который позволяет строить рекомендации по подбору журнала для осуществления публикации по тексту рукописи, предлагаемой к публикации [6].

Проект Google Scholar использует двухуровневый классификатор с 8 элементами первого уровня и 400 элементами второго уровня. Разбиение по темам может использоваться на странице scholar.google.com/citations для фильтрации журналов при показе их показателей (h5-индекса и h5-медианы), что позволяет строить более объективные рейтинги для каждой из тематических областей, заданных в классификаторе (Рис. 1). Тематическая фильтрация возможна только для англоязычных журналов. Для русскоязычных журналов, как и для журналов, издаваемых на других языках, тематическая классификация отсутствует.

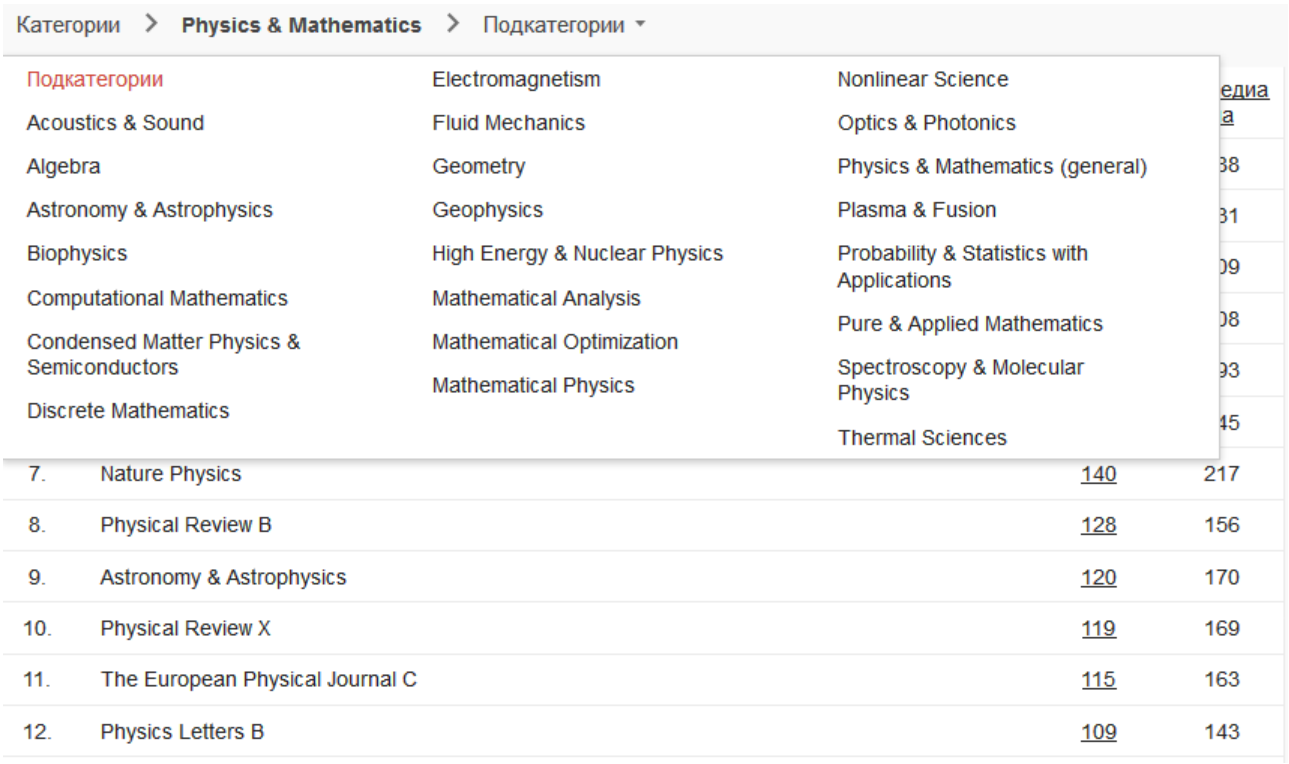


Рис. 1. Интерфейс тематической фильтрации при оценке журнала.

Для корректировки данных Google, в соответствии со своей основной методикой, активно использует взаимодействие с пользователем для сбора информации «снизу–вверх», позволяя авторам создавать собственные страницы со списком статей, фотографией, описаниями интересов (Google Scholar Citations). Добавление статей в профили может производиться автоматизированно (пользователю предлагаются подобранные варианты) или вручную с указанием полных библиографических данных.

Проект Scopus использует двухуровневый классификатор All Science Journal Classification Codes (ASJC), содержащий 4 записи первого уровня и около 350 записей второго уровня для классификации журналов. На странице www.scival.com можно посмотреть распределение журналов по областям и относительные нормированные характеристики для выбранных разделов классификатора, изменение количества публикаций по тематикам с течением времени и другие показатели. Данные доступны при наличии платной подписки.

Проект РИНЦ использует трехуровневый Государственный Рубрикатор НТИ России (ГРНТИ), содержащий около 8 тыс. рубрик (elibrary.ru/rubrics.asp). Тематическую классификацию можно использовать для поиска журналов и статей, а

также для фильтрации результатов отбора журналов при выдаче их наукометрических параметров.

Проект Open Academic Graph (OAG), являющийся расширенной версией Microsoft Academic Graph (MAG), содержит 170 млн. статей со ссылками цитирования. Проект не является наукометрической системой или системой цитирования, однако данные проекта могут использоваться для апробации алгоритмов наукометрических систем. Данные можно свободно скачать с сайта проекта www.aminer.org/open-academic-graph.

Кроме перечисленных выше классификаторов коммерческих систем существует целый ряд общепринятых классификаторов, не связанных с какой-то конкретной системой цитирования. На мировом уровне наиболее известным считается трехуровневый классификатор OECD Fields of Science, содержащий более двухсот рубрик, который планировалось использовать, в том числе, в проекте «Карта российской науки». Во многих российских журналах для тематической классификации статей самими авторами используется более подробная Универсальная десятичная классификация (УДК), содержащая более 150 тысяч рубрик. Также для тематической классификации различных научных материалов используются Рубрикатор ВИНИТИ, содержащий более 53 тысяч рубрик, и целый ряд других тематических классификаторов: Классификатор Российского научного фонда (РНФ) [7]; Классификатор Российского фонда фундаментальных исследований(РФФИ) [8]; Международная патентная классификация (МПК) [9], Общероссийский классификатор стандартов (ОКС) [10], Mathematics Subject Classification (MSC) [11]; Journal of Economic Literature Classification (JEL) [12] и другие (scs.viniti.ru/MapService/treeList.aspx). При наличии такого многообразия классификаторов закономерным является появление различных проектов по их согласованию, например, проект по сопоставлению классификаторов Scopus и OECD [13], а также проект ВИНИТИ [14].

Перечисленные выше проекты ставили своей целью разработку систем подсчета показателей цитирования научных публикаций и проведение их тематической классификации по областям науки. Следующим шагом развития стало появление на их основе систем оценки научной деятельности организаций в целом.

Испанский проект SCImago Journal & Country Rank Гранадского университета (или «Атлас науки») оценивает на основе данных Scopus агрегированные данные

о научной деятельности в Испании, Португалии и странах Южной Америки. На сайте проекта www.scimagojr.com приводятся показатели не только по научным журналам, но и по странам в целом. Индекс SJR, разработанный авторами проекта, является альтернативой импакт-фактору.

Проект Faculty Scholarly Productivity Index (FSPI) оценивает на основе данных Scopus метрические показатели университетов США. Помимо количества публикаций и показателей цитирования в этом проекте для расчета ранга университета используются данные о полученных наградах и премиях, а также об объемах федерального финансирования исследований. На основе агрегированных данных производится ранжирование более чем 350 университетов.

Проект Times Higher Education (THE) ставит своей задачей оценку университетов всего мира [15]. Разработанный в рамках проекта индекс World University Rankings строится на основе данных о цитируемости WoS, которые составляют 32.5% рейтинга [16]. Помимо этого, учитываются субъективные оценки экспертов, объем финансирования проведенных исследований, привлечение иностранных студентов и преподавателей, а также внедрение разработок вуза в промышленность.

Проект QS World University Rankings [17] оценивает по показателям исследовательской и преподавательской деятельности, соотношению студентов и преподавателей, среднему индексу цитирования в расчете на одного преподавателя, репутации у работодателей, а также количеству иностранных студентов и преподавателей.

Проект Academic Ranking of World Universities (ARWU), который часто называют «Шанхайским рейтингом» (www.shanghairanking.com), учитывает получение выпускниками университета Нобелевских премий, количество опубликованных статей в журналах Nature и Science и показателей цитируемости.

Следует отметить, что проведение подобных сравнений без учета языка преподавания, так же, как и оценка журналов без учета их тематической области, дает не вполне точные результаты [18]. Например, сравнение университетов всего мира по уровню цитируемости только в англоязычных журналах неоспоримо показывает только тот факт, что процент преподавателей и студентов, свободно владеющих английским языком в университетах США, Англии и Канады,

значительно выше, чем в России или других не англоязычных странах. Аналогично, сравнение доли иностранных студентов и преподавателей в университетах с английским и японским языками преподавания показывает не столько уровень образования в учебном заведении, сколько количество иностранцев, свободно владеющих данным языком.

Для наукометрических систем, которые ставят своей задачей получение объективных и сбалансированных оценок качества научной продукции, учет области проведения исследования при анализе наукометрических данных, в том числе языка, тематической области и других подобных характеристик, является необходимым требованием при построении объективных наукометрических показателей.

2. ТЕМАТИЧЕСКИЙ АНАЛИЗ С ИСПОЛЬЗОВАНИЕМ ТЕКСТОВОЙ ИНФОРМАЦИИ И КЛАССИФИКАТОРОВ

В процессе разработки и развития наукометрической системы ИСТИНА особое внимание всегда уделялось развитию методов интеллектуального анализа информации, в том числе, методам тематического анализа. Объем данных, обрабатываемых этой системой в настоящее время, значительно уступает мировым системам цитирования, поскольку охватывает всего 28 организаций, 900 тысяч публикаций, 70 тысяч монографий и 13 тысяч патентов. Однако количество типов используемых данных значительно выше. Кроме публикаций и патентов в системе присутствует полная информация о данных по научным проектам (НИР, НИОКР, гранты), докладах на конференциях, диссертациях и дипломах, об участии сотрудников в деятельности различных советов и редколлегиях, получаемых ими премиях и наградах, читаемых учебных курсах и других данных [19]. Кроме того, информация в системе проходит двойную проверку. Основным принципом работы системы является движение информации «снизу-вверх». На первом этапе пользователь, как наиболее заинтересованное лицо, регистрирует в системе все свои работы, которые отображаются на его персональной странице. На втором этапе ответственные сотрудники подразделений подтверждают достоверность данных. Подобный метод сбора информации с использованием создания персональных страниц в настоящий момент использует в проекте Google Scholar Citations корпорация Google, являющаяся лидером на рынке обработки текстовых

данных. Но в силу объективных причин в этой системе невозможно организовать второй этап верификации.

Одним из наиболее простых способов проведения тематического анализа является использование классификаторов с ручным сопоставлением объектов и тематических классов, в том числе использование тематической классификации карточек журналов. Такой подход используется в Scopus, WoS и РИНЦ.

В наукометрической системе ИСТИНА на начальном этапе для анализа активности сотрудников и организаций на различных тематических направлениях также были реализованы методы анализа с использованием рубрикации журналов по различным статическим рубрикаторам. С использованием интерактивного интерфейса на странице статистики организации [20] представлены данные о распределении числа статей, цитируемости WoS, числа авторов и других агрегированных характеристик по рубрикам Scopus и ГРНТИ. Данные могут предоставляться как по отдельным подразделениям, так и по организации в целом с возможностью фильтрации по году публикации, преодолению порогового значения анализируемого показателя и принадлежности к группе журналов: журналы из Scopus, журналы из Top25, журналы из ВАК, сборники статей и другие. При этом можно отдельно указывать метрику фильтрации по пороговому значению и метрику для отображения на диаграмме. Например, можно производить фильтрацию по количеству статей, отображать число ссылок на статью.

Возможно проведение анализа как на уровне организации в целом, так и на уровне каждого подразделения отдельно. Следует отметить, что выбор уровня агрегации особенно полезен с учетом неоднозначности определения подразделения для каждой отдельно взятой публикации. В крупных научных организациях большое количество статей публикуется в соавторстве сотрудниками разных подразделений. При традиционном способе подсчета агрегированные данные по отдельным подразделениям считаются независимо, а потом суммируются. В этом случае совместные статьи учитываются несколько раз, что приводит к искажению итоговых показателей. Использование возможности агрегирования исходных данных как на уровне организации (Рис. 2), так и на уровне подразделения (Рис. 3) делает такие оценки более точными и объективными.

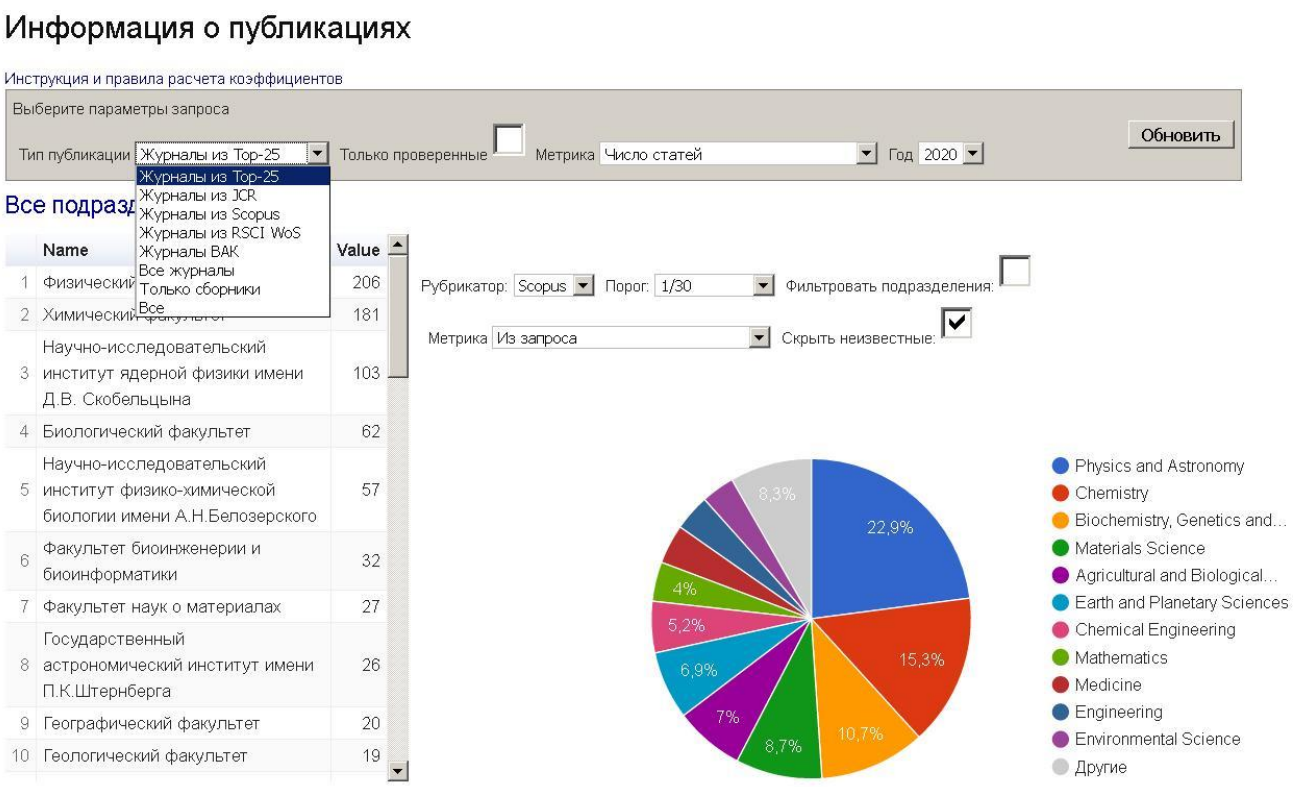


Рис. 2. Интерфейс для анализа распределения числа статей по рубрикам организации.

Подобный подход предоставляет пользователю возможность оценить степень публикационной активности сотрудников в различных тематических областях. Однако он не позволяет анализировать информацию с достаточной степенью детализации. Рубрикатор является статическим, и дополнительная детализация внутри одной рубрики невозможна.

Вторым возможным подходом являются определение тематики и поиск информации по ключевым словам, аннотациям или полным текстам статей. Ключевые слова могут задаваться авторами работы при ее регистрации в системе или вычисляться в процессе индексации из аннотации, полных текстов статей или списка цитируемой литературы, например, Author Keywords и KeyWords Plus в WoS. Такой подход позволяет в большей степени конкретизировать тематику поиска, которая необходима для таких задач, как выделение новых тематических направлений или поиск информации по конкретной информационной потребности пользователя. Следует отметить, что применение подобного тематического

анализа не ограничивается только поиском информации. Например, в работе [21] предлагается использовать тематический анализ для оценки качества журнала.

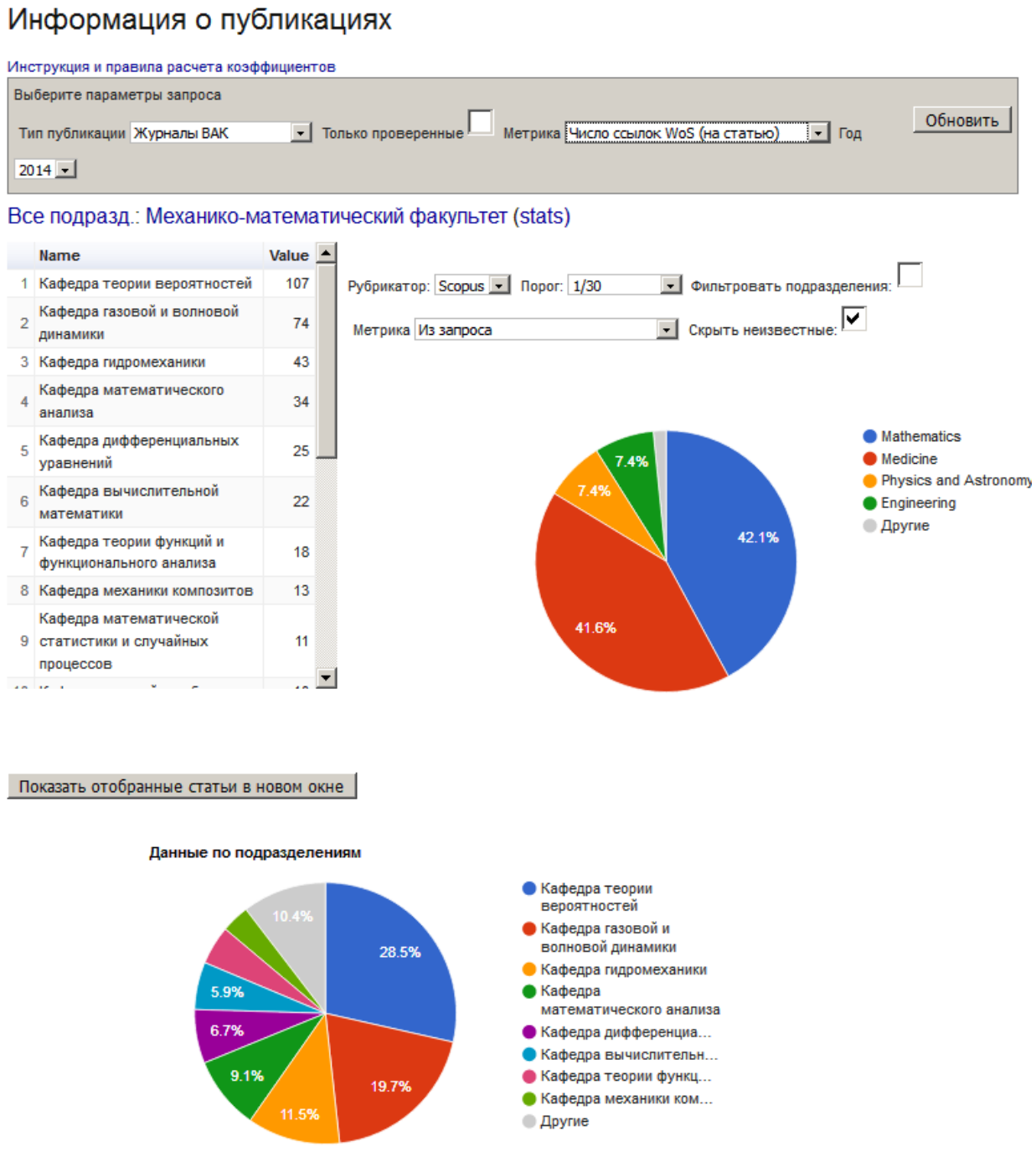


Рис. 3. Интерфейс для анализа распределения данных о цитировании статей подразделения по разным тематическим рубрикам.

Основная гипотеза состоит в том, что в «хороших» научных журналах статьи должны быть посвящены фиксированному набору тематик и эти тематики

должны меняться со временем. Таким образом, после обучения алгоритма тематического анализа на обучающем наборе статей из всех анализируемых журналов можно провести тематическую и временную классификацию статей из этих журналов. Качество журнала будет пропорционально точности классификации, с которой содержащиеся в нем статьи были правильно классифицированы по принадлежности к журналу и к временному промежутку публикации.

Основная сложность использования для тематического анализа ключевых слов состоит в ограниченности его набора. При описании статей авторы обычно указывают менее 10 слов. Например, среднее количество ключевых слов, которое авторы указывают при регистрации статей в ИАС ИСТИНА, составляет 3.8. Дополнительным препятствием является субъективность выбора. На первом этапе авторы выделяют из статьи основные понятия, которые, по их оценке, являются значимыми на данный момент. На втором этапе для каждого понятия указывается только одно его отображение на множество ключевых слов без учета возможных синонимов. Таким образом, статьи схожей тематики могут иметь непересекающийся набор ключевых слов, и точность определения их тематического сходства значительно снижается. Вместе с тем, подобный подход к поиску схожих по тематике статей реализован в некоторых системах цитирования. Например, протестировать качество подбора статей при осуществлении поиска по ключевым словам в русскоязычных журналах можно на поисковой странице проекта РИНЦ [22].

Проект WoS предоставляет пользователям сервис Manuscript Matcher подбора журнала для публикации по тексту статьи. Для работы сервис требует предварительной регистрации пользователя. После загрузки заголовка статьи и ее аннотации сервис определяет ключевые слова и ищет соответствие с ключевыми словами журналов. Результат показывается в виде списка журналов с указанием описания журнала, а также списка общих ключевых слов с указанием меры сходства с загруженной статьей. Сервис может быть полезен для авторов, которые используют узкоспециализированные термины, например, в химии, биологии или астрономии. Для более общих тем сопоставление терминов дает не очень точный результат. Например, для статьи "Determining the thematic proximity of scientific journals and conferences using Big Data technologies" лучшими журналами в результатах поиска являются "Scientometrics" и "Journal of the association for information

science and technology", однако в top5 попадают журналы "Journal of medical systems" и "Journal of digital imaging" по терминам "create software tools" и "full-text information".

Проект РИНЦ предлагает пользователям сервис поиска похожих статей. Пользователь может выбрать одну из статей, уже проиндексированных в системе, и запросить поиск похожих статей по тематике. Но результаты такого поиска обладают еще меньшей точностью, чем результаты поиска по ключевым словам и работы сервиса Manuscript Matcher. Например, для статьи «Архитектура, методы и средства базовой составляющей системы управления научной информацией «ИСТИНА-НАУКА МГУ»» определяется 14 тысяч близких по тематике статей, и в списке top10 нет ни одной статьи, которая была бы связана с системой, рассматриваемой в статье, или каким-либо аналогом, и только одна статья затрагивает вопросы наукометрии. В результатах поиска по тематическому сходству top3 составляют: "Information technology of software architecture structural synthesis of information system", «Анализ применения asp.net при разработке информационной системы ‘Analysis of the asp.net development information system’», «Общий обзор agris (agricultural research information system)».

Одним из возможных способов улучшения полноты поиска по ключевым словам и разрешения вопросов омонимии явля.тся расширение набора ключевых слов на основе построения связей между ключевыми словами [23], а также использование переводов терминов. В проекте ИСТИНА для автоматизации процесса перевода используются материалы Википедии, а также бесплатные сервисы компании Abbyy. Поиск по ключевым словам используется в качестве первого этапа тематического анализа в разрабатываемых алгоритмах поиска экспертов и подбора журналов, которые апробируются в настоящее время на данных системы ИСТИНА. Результаты исследований, проводимых в этом направлении, нельзя еще использовать для реализации промышленного ПО, однако уже сейчас можно утверждать, что использование только тематического анализа на основе ключевых слов, аннотаций и текстов не позволяет получить удовлетворительный результат классификации. В этой связи в разрабатываемых алгоритмах использу-

ется комбинирование методов полнотекстового анализа и методов анализа теории графов, которые анализируют явные или неявные связи между классифицируемыми объектами.

3. ТЕМАТИЧЕСКИЙ АНАЛИЗ С ИСПОЛЬЗОВАНИЕМ СВЯЗЕЙ ОБЪЕКТОВ

Использование связей между объектами (или граф объектов) позволяет дополнить или уточнить данные анализа в случае недостатка информации. Объекты в графе могут быть одного типа, например, статьи и ссылки между статьями, или иметь разный тип, например, сотрудники и их проекты. Целью анализа графа могут быть расширение области поиска или уточнение значимости объектов в существующей области поиска.

Одним из примеров дополнения данных в графе с объектами разного типа является задача поиска экспертов по заданной тематике [24]. Для поиска экспертов в графе авторства определяются объекты, наиболее связанные с экспертами (статьи, монографии, проекты, отчеты и другие), а также степень связи, выделяются ключевые слова объектов, на основе расширенного набора ключевых слов и весов связей графа строится информационный портрет пользователя и оценивается его близость с исходным поисковым запросом. Для апробации алгоритма использовались данные наукометрической системы ИСТИНА. Применение подобных алгоритмов в системах подсчета цитирования затруднено, поскольку граф связей объектов в них содержит только два типа вершин: авторы и публикации. В полноценных наукометрических системах информационный портрет пользователя составляется из большего количества типов объектов, что улучшает качество результатов.

Примером решения задачи уточнения данных на основе связей между аналогичными объектами является алгоритм определения авторства статьей [25], который реализован в системе ИСТИНА. Предполагается, что авторские коллективы обладают определенной устойчивостью, и вероятность публикации двумя авторами нескольких совместных статей гораздо выше, чем написание статьи, в которых одного из авторов заменяет полный однофамилец. В соответствии с этой гипотезой для разрешения неоднозначности при определении авторов статьи среди всех возможных однофамильцев строится граф соавторства и выбирается наиболее связанная компонента.

Используя граф соавторства, можно также решать задачу определения тематической близости журналов без использования данных полнотекстового анализа. Основной гипотезой при реализации этого метода является предположение, что значительная часть авторов публикует статьи в своей предметной области и, следовательно, несколько журналов, в которых публикуется одинаковый набор авторов, близки по тематике. Исходя из этой гипотезы, тематическая близость двух журналов рассчитывается как взвешенная сумма авторов, имеющих публикации в обоих журналах. При этом учитывается не только количество публикаций, сделанных автором, но и позиция автора в библиографическом описании статьи. Вес связи статьи распределяется по всем авторам, но первые авторы имеют больший вес, чем остальные. Формальное описание алгоритма приводится в работе [26]. Основным отличием данного алгоритма от аналогичных алгоритмов, использующих полнотекстовый анализ или анализ по ключевым словам, является нечувствительность к языку журналов и, как следствие, возможность поиска связей журналов на разных языках. Кроме того, алгоритм не нуждается в длительном обучении на больших массивах текстов, показывая при этом достаточно высокую точность 78%.

Дальнейшим развитием описанного в работе [26] алгоритма явились работы по автоматизации расширения области поиска журналов в графе соавторства. Основной предпосылкой является предположение о транзитивности соотношения близости для узкоспециализированных журналов. Если два узкоспециализированных журнала близки по тематике третьему, то они близки между собой. Вместе с тем, обобщение этого правила на все, в том числе общетематические, журналы является неверным. Например, наличие общих авторов у каких либо двух журналов с общетематическим изданием «Известия РАН» не означает взаимной тематической близости исходных двух журналов. В этой связи необходимо использовать математические модели с нормированием весов ребер графа связей журналов [26]. В ходе проведенных исследований было показано, что наилучший результат достигается при проведении нормировки весов ребер с использованием общей суммы исходящих из каждой вершины ребер. После проведения нормировки матрица близости между журналами считается на основе сравнения путей в графе длинной 3. Подобный подход позволяет значительно увеличить

полноту тематического поиска. Окончательный результат строится на основе объединения двух списков: наиболее близкие журналы в исходной матрице тематической близости и наиболее близкие журнал в расширенной матрице тематической близости. Объединение этих списков перед показом пользователю позволяет увеличить полноту поиска, не сильно снижая его точность.

Программная реализация алгоритма используется в системе ИСТИНА для предоставления пользователям удобного интерфейса тематического поиска журналов. Для осуществления поиска пользователю необходимо выбрать один известный ему журнал по заданной тематике, найдя его по названию (Рис. 4).

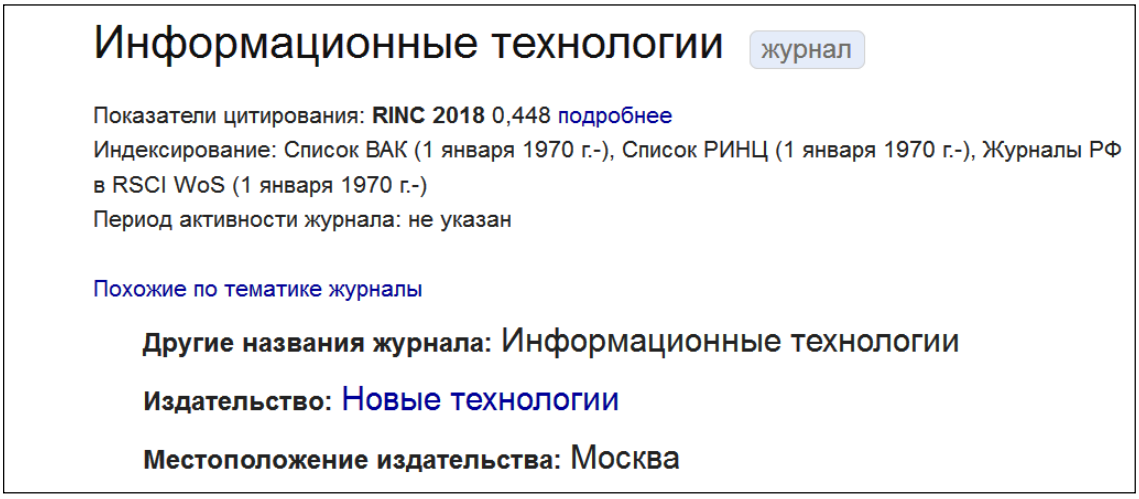


Рис. 4. Карточка журнала.

После этого нужно перейти по ссылке «Похожие по тематике журналы». Для удобства работы в строке каждого журнала в представленном списке указываются оценка его тематического сходства с исходным журналом, различные характеристики цитируемости и количество статей из этого журнала, загруженных в наукометрическую систему (Рис. 5).

Пользователь может перейти на страницу журнала или продолжить перемещение по графу тематических связей журналов, используя ссылки в столбце «Похожие журналы».

Следует отметить, что данный алгоритм может искать тематические связи не только между журналами, но и между другими группами объектов, имеющих авторов. В данном примере алгоритм также ищет конференции, похожие по тематике на заданный журнал.

Show by 10 items Search:

N	Журнал	Вес	Статей за 5 лет	WS	SJR	RINC	Похожие журналы	Похожие конференции
1	Доклады Академии наук	40,7	1061	.195 (1999)	-	1.058 (2018)	журналы	конференции
2	Программная инженерия	39,07	56	-	-	.353 (2017)	журналы	конференции
3	Наука и образование (МГТУ им. Н.Э. Баумана) (электронный журнал)	35,3	17	-	-	-	журналы	конференции
4	Нейрокомпьютеры: разработка, применение	33,38	42	-	-	.341 (2017)	журналы	конференции
5	Программирование	26,62	51	-	-	.685 (2018)	журналы	конференции
6	Programming and Computer Software	21,57	76	.637 (2019)	-	-	журналы	конференции
7	Проблемы информатики	19,27	1	-	-	.138 (2017)	журналы	конференции

Рис. 5. Интерфейс тематической фильтрации при оценке журнала.

Еще одной важной практической задачей, которая может решаться с использованием описания связей между объектами в наукометрической системе, является задача определения авторитетности экспертов при осуществлении их поиска по тематическому описанию. Для ориентированных графов классическим алгоритмом оценки авторитетности вершин в графе является алгоритм PageRank, который использовался в системе Google для ранжирования результатов поиска. Алгоритм основан на предположении, что входящее ребро в графе подтверждает авторитетность вершины, причем значимость этого подтверждения тем выше, чем выше авторитетность исходящей вершины. В наукометрических системах алгоритм может эффективно использоваться для анализа графа цитируемости. Для анализа неориентированного графа соавторства и других подобных графов в наукометрических системах возможно использование целого ряда других характеристик: степень связности (количество ребер для каждой вершины); степень близости (среднее кратчайшее расстояние до других вершин графа); степень посредничества (количество кратчайших путей между всеми парами вершин, проходящих через заданную вершину); степень влиятельности (степень связности, в

которой вклад каждого ребра зависит от степени влиятельности соседней вершины, например, PageRank); кросс-кликовая центральность (число клик, которым принадлежит узел) и другие. Предварительные эксперименты, проведенные на данных системы ИСТИНА, показали, что такой подход может быть достаточно эффективен для использования при ранжировании результата поиска экспертов, автоматического определения устойчивых научных коллективов и других подобных задач.

ЗАКЛЮЧЕНИЕ

Использование алгоритмов тематического анализа для решения целого ряда задач обработки информации в наукометрических системах позволяет создавать удобные сервисы для поиска и обработки информации. Комбинирование полнотекстовых и графовых методов анализа позволяет увеличить точность и полноту представляемых результатов. В настоящий момент в системах научного цитирования такие сервисы представлены недостаточно широко. Научные изыскания на этом направлении, проводимые с использованием данных проекта ИСТИНА, могут позволить получить новые механизмы поиска и обработки наукометрической информации.

Благодарности

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект 18-07-01055.

СПИСОК ЛИТЕРАТУРЫ

1. *Акоев М.А., Маркусова В.А., Москалева О.В., Писляков В.В.* Руководство по наукометрии: индикаторы развития науки и технологии. Екатеринбург: Издательство Уральского университета, 2014. 248 с.

2. *Орлов А.И.* Наукометрия и управление научной деятельностью // Управление большими системами. Специальный выпуск 44: Наукометрия и экспертиза в управлении наукой. Институт проблем управления им. В.А. Трапезникова РАН. 2013. С. 538–568.

3. *Бричковский В.В.* Наукометрический анализ в информационном обеспечении инновационной деятельности // В мире науки. 2017. № 8(174). С. 64–67.

4. *Афонин С.А., Козицын А.С., Шачнев Д.А.* Программные механизмы агрегации данных, основанные на онтологическом представлении структуры реляционной базы наукометрических данных // Программная инженерия. 2016. Т. 7, №9. С. 408–413.

5. *Afonin S.* Ontology models for access control systems // Proc. of the 3rd International Conference Russian-Pacific Conference on Computer Technology and Applications (RPC). 2018. P. 1–6.

6. Сервис подбора журнала WoS. URL: http://mjl.clarivate.com/home

7. Классификатор РНФ. URL: http://www.rscf.ru/node

8. Классификатор РФФИ. URL: http://www.rfbr.ru/rffi/ru/contest_documents

9. Классификатор МПК. URL: http://www.fips.ru

10. Классификатор ОКС. URL: http://classinform.ru/oks.html

11. Классификатор MSC. URL: http://www.ams.org/msc/

12. Классификатор JEL.
URL: http://www.aeaweb.org/journal/jel_class_system.html

13. Проект по сопоставлению классификаторов Scopus и OECD.
URL: http://report03.metrics.ekt.gr/en/appendixIII

14. Проект по сопоставлению классификаторов ВИНИТИ.
URL: http://scs.viniti.ru/ MapService/mapform.aspx

15. Проект Times Higher Education.
URL: http://www.timeshighereducation.com

16. Индекс World University Rankings.
URL: http://gtmarket.ru/ratings/the-world-university-rankings/info

17. Проект QS World University Rankings. URL: http://www.topuniversities.com

18. *Кинчарова А.В.* Методология мировых рейтингов университетов: анализ и критика // Университетское управление: практика и анализ. 2014. № 2. С. 70–80.

19. Данные проекта ИСТИНА. URL: http://istina.msu.ru/statistics/activity/

20. Статистика организации в проекте ИСТИНА.
URL: http://istina.msu.ru/ statistics/organization/214524/dynamic

21. *Краснов Ф.В.* Сравнительный анализ коллекций научных журналов // Труды СПИИРАН. 2019. Т. 18. С. 767–793.

22. Поиск по ключевым словам в системе РИНЦ.

URL: https://www.elibrary.ru/ querybox.asp

23. *Афонин С.А., Лунев К.В.* Выявление тематических направлений в коллекции наборов ключевых слов // Программная инженерия. 2015. № 2. С. 29–39.

24. *Vasenin V., Lunev K., Afonin S., Shachnev D*. Methods for intelligent data analysis based on keywords and implicit relations: The case of "ISTINA" data analysis system. In Proc. of the International Conference Actual Problems of Systems and Software Engineering (APSSE 2019), IEEE Conference Proceedings, P. 151–155, US, 2019.

25. *Козицын А.С., Афонин С.А.* Разрешение неоднозначностей при определении авторов публикации с использование графов соавторства в больших коллекциях библиографических данных // Программная инженерия. 2017. Т. 8, № 12. С. 556–562.

26. *Козицын А.С., Афонин С.А.* Нахождение скрытых зависимостей между объектами на основе анализа больших массивов библиографических данных // In Proc. of the International Conference Actual Problems of Systems and Software Engineering (APSSE 2019), IEEE Conference Proceedings. 2019. P. 320–328.

THE USE OF THEMATIC ANALYSIS METHODS IN SCIENTOMETRIC SYSTEMS

A. S. Kozitsin, S. A. Afonin, D. A. Shachnev

Institute of Mechanics Lomonosov Moscow State University

alexanderkz@mail.ru, serg@msu.ru, mitya57@gmail.com

Abstract

Modern scientometric systems and citation systems use various mechanisms of thematic search and thematic filtering of information. In most cases, a full-text approach is used for thematic analysis of articles and journals, which has a number of limitations. The use of algorithms based on graph analysis, both independently and in conjunction with full-text algorithms, eliminates these limitations and improves the completeness and accuracy of subject search. The algorithm developed by the authors and presented in this work uses the co-authorship graph to analyze the thematic proximity of journals. The algorithm is insensitive to the language of the journal and selects similar journals in different languages, which is difficult to implement for algorithms

based on the analysis of full-text information. The algorithm was tested in the scientometric system IAS ISTINA. In the interface developed for these purposes, the user can select one journal that is close to him on the subject, and the system will automatically generate a selection of journals that may be of interest to the user both in terms of studying the materials available in them and in terms of publishing his own articles. In the future, the developed algorithm can be adapted to search for similar conferences, collections of publications and scientific projects. The presence of such a tool will increase the publication activity of young employees, increase the citation rate of articles and the citation rate between journals. The results of the algorithm for determining thematic proximity between journals, collections, conferences and scientific projects can also be used to build rules in models of differentiating access to data based on domain ontologies.

Keywords: *thematic classification, bibliographic data, co-authorship graph, information systems.*

REFERENCES

1. *Akoev M.A., Markusova V.A., Moskaleva O.V., Pisliakov V.V.* Rukovodstvo po naukometrii: indikatory razvitiia nauki i tekhnologii. Ekaterinburg: Izdatelstvo Uralskogo universiteta, 2014. 248 s.

2. *Orlov A.I.* Naukometriia i upravlenie nauchnoi deiatelnostiu // Upravlenie bolshimi sistemami. Spetsialnyi vypusk 44: Naukometriia i ekspertiza v upravlenii naukoi. Institut problem upravleniia im. V. A. Trapeznikova RAN. 2013. S. 538–568.

3. *Brichkovskii V.V.* Naukometricheskii analiz v informatsionnom obespechenii innovatsionnoi deiatelnosti // V mire nauki. 2017. № 8(174). S. 64–67.

4. *Afonin S.A., Kozitsyn A.S., Shachnev D.A.* Programmnye mekhanizmy agregatsii dannykh, osnovannye na ontologicheskom predstavlenii struktury reliatsionnoi bazy naukometricheskikh dannykh // Programmnaia inzheneriia. 2016. T. 7, №9. S. 408–413.

5. *Afonin S.* Ontology models for access control systems // Proc. of the 3rd International Conference Russian-Pacific Conference on Computer Technology and Applications (RPC). 2018. P. 1–6.

6. Servis podbora zhurnala WoS. URL: http://mjl.clarivate.com/home

7. Klassifikator RNF. URL: http://www.rscf.ru/node

8. Klassifikator RFFI. URL: http://www.rfbr.ru/rffi/ru/contest_documents

9. Klassifikator MPK. URL: http://www.fips.ru

10. Klassifikator OKS. URL: http://classinform.ru/oks.html

11. Klassifikator MSC. http://www.ams.org/msc/

12. Klassifikator JEL.
URL: http://www.aeaweb.org/journal/jel_class_system.html

13. Proekt po sopostavleniiu klassifikatorov Scopus i OECD.
URL: http://report03. metrics.ekt.gr/en/appendixIII

14. Proekt po sopostavleniiu klassifikatorov VINITI.
URL: http://scs.viniti.ru/ MapService/mapform.aspx

15. Proekt Times Higher Education.
URL: http://www.timeshighereducation.com

16. Indeks World University Rankings.
URL: http://gtmarket.ru/ratings/the-world-university-rankings/info

17. Proekt QS World University Rankings. URL: http://www.topuniversities.com

18. *Kincharova A.V.* Metodologiia mirovykh reitingov universitetov: analiz i kritika // Universitetskoe upravlenie: praktika i analiz. 2014. No. 2. S. 70–80.

19. Dannye proekta ISTINA. URL: http://istina.msu.ru/statistics/activity/

20. Statistika organizatsii v proekte ISTINA.
URL: http://istina.msu.ru/statistics/ organization/214524/dynamic

21. *Krasnov F.V.* Sravnitelnyi analiz kollektsii nauchnykh zhurnalov // Trudy SPIIRAN. 2019. T. 18. S. 767–793.

22. Poisk po kliuchevym slovam v sisteme RINTs.
URL: https://www.elibrary.ru/querybox.asp

23. *Afonin S.A., Lunev K.V.* Vyiavlenie tematicheskikh napravlenii v kollektsii naborov kliuchevykh slov // Programmnaia inzheneriia. 2015. № 2. S. 29–39.

24. *Vasenin V., Lunev K., Afonin S., Shachnev D*. Methods for intelligent data analysis based on keywords and implicit relations: The case of "ISTINA" data analysis system // In Proc. of the International Conference Actual Problems of Systems and Software Engineering (APSSE 2019), IEEE Conference Proceedings, 2019. P. 151–155, US, 2019.

25. *Kozitsyn A.S., Afonin S.A.* Razreshenie neodnoznachnostei pri opredelenii avtorov publikatsii s ispolzovanie grafov soavtorstva v bolshikh kollektsiiakh bibliograficheskikh dannykh // Programmnaia inzheneriia. 2017. T. 8, No 12. S. 556–562.

26. *Kozitsyn A.S., Afonin S.A.* Nakhozhdenie skrytykh zavisimostei mezhdu obieektami na osnove analiza bolshikh massivov bibliograficheskikh dannykh // In Proc. of the International Conference Actual Problems of Systems and Software Engineering (APSSE 2019), IEEE Conference Proceedings. 2019. P. 320–328.

СВЕДЕНИЯ ОБ АВТОРАХ

КОЗИЦЫН Александр Сергеевич – ведущий научный сотрудник, к. ф.-м. н., окончил мехмат МГУ им. М.В. Ломоносова. Специалист в области информационного поиска и баз данных.

Alexander Sergeevich KOZITSYN – Leading Researcher, Ph.D., graduated from M.V. Lomonosov Moscow State University. Specialist in the field of information retrieval and database.

email: alexanderkz@mail.ru,
ORCID: 0000-0002-8065-9061

АФОНИН Сергей Александрович – ведущий научный сотрудник, к. ф.-м. н., окончил мехмат МГУ им. М.В. Ломоносова. Специалист в области регулярных языков и информационных систем.

Sergey Alexandrovich AFONIN – Leading Researcher, Ph.D., graduated from M.V. Lomonosov Moscow State University. Specialist in the field of regular languages and information systems.

email: serg@msu.ru, ORCID:0000-0003-3058-9269

Шачнев Дмитрий Алексеевич – программист, окончил мехмат МГУ им. М.В. Ломоносова. Специалист в области информационных систем.

Dmitry Alekseevich SHACHNEV – programmer, graduated from M.V. Lomonosov Moscow State University. Specialist in information systems.

email: mitya57@gmail.com, ORCID: 0000-0002-5940-9180

Материал поступил в редакцию 18 ноября 2020 года

УДК 004.04+004.9

ИССЛЕДОВАНИЕ КОНТЕКСТОВ ЭКОСИСТЕМЫ «ЦИФРОВОГО ТУРИЗМА»

О. В. Кононова[1, 4, [0000-0001-6293-7243]], **Д. Е. Прокудин**[1, 2, 4, [0000-0002-9464-8371]], **Е. Н. Тупикина**[3, 4, [0000-0001-9531-9900]]

[1] *Университет ИТМО, Санкт-Петербург, Россия*

[2] *Санкт-Петербургский государственный университет, Санкт-Петербург, Россия*

[3] *Дальневосточный федеральный университет, Владивосток, Россия*

[4] *Центр исследований цифрового общества, Россия*

[1]kononolg@yandex.ru, [2]hogben.young@gmail.com, [3]etupikina@mail.ru

Аннотация

Современные информационно-коммуникационные технологии, элементы цифровизации постоянно и стремительно развиваются, что, в свою очередь, оказывает непосредственное влияние на все сферы человеческой деятельности. В свете последних событий, связанных с коллапсом туристического бизнеса из-за COVID-19, большой научный интерес проявляется к сфере услуг, а именно, к сфере «цифрового туризма». Цифровой туризм опирается на широкое внедрение новых технологий, таких как социальные сети и мобильные технологии, умные устройства и датчики для сбора и использования огромного количества данных для создания новых ценностных предложений. В связи с этим авторами поставлена цель – представить обзор литературы по «цифровому туризму» с позиций научного и медиа дискурса. Авторами представлен комплексный науковедческий подход, включающий последовательное выполнение всех этапов обзора от определения терминологического ядра междисциплинарного направления, формирования поисковых запросов, каскадного поиска, подбора и контент-анализа материалов до выявления и экспликация контекстов. Источниками информации для подготовки обзора выступили публикации из академических баз данных: Web of Science, Science-Direct, Scopus, GoogleScholar, eLibrary, Киберленинка, а также материалы и публикации в русскоязычных СМИ – Интегрум.

Полученные результаты будут полезны ученым при определении перспективных направлений исследований в области «цифрового туризма», а также позволят углубить знания о механизмах поиска, сбора и анализа данных и интегрированных и аналитических средах.

***Ключевые слова**: информационно-коммуникационные технологии, цифровые трансформации, цифровой туризм, электронный туризм, eTourism, smart tourism.*

ВВЕДЕНИЕ

Стремительное развитие информатики, интернета, информационно-коммуникационных технологий накладывает отпечаток на социальное развитие общества и экономический вектор развития. В век цифровизации создаются, сохраняются, записываются и накапливаются огромные массивы структурированной и неструктурированной информации, формируя большие данные. Все это заставляет компании думать о перспективах и создавать новые бизнес-модели.

Туризм является одной из важнейших составляющих экономики многих стран мира, которая обеспечивает рабочие места населению, загрузку гостиниц, отелей, ресторанов, а также поступление иностранной валюты.

На фоне новой короновирусной пандемии 2020 года сфера туризма оказалась одной из наиболее уязвимой. Этот факт и закрепил предположения авторов об актуальности и своевременности предлагаемого исследования.

Следует отметить, что для Российской Федерации туризм является одной из отраслей экономики, которая одновременно играет социальную и экономическую роли. В связи с этим в конце 2019 года Правительство РФ утвердило «Стратегию развития туризма в Российской Федерации на период до 2035 года», где отражены первоочередные задачи туризма: «… достижение уровня мировых лидеров в развитии цифровой инфраструктуры и сервисов, развитие цифровых платформ продвижения туристских продуктов и брендов, цифровых средств навигации и формирования туристского продукта» [1].

Стратегия предлагает решить две важные задачи – создать конкурентоспособный туристический продукт и сделать его востребованным и доступным. Для

этого предполагаются комплексное развитие и благоустройство туристских территорий и инфраструктуры, цифровизация индустрии туризма, упрощение визовых формальностей и многое другое. Одной из главных задач, представленной в стратегии, является трансформация туристкой отрасли РФ на базе цифровых технологий. В частности, речь идет о развитии цифровых платформ, которые предоставляют большой выбор услуг сферы гостеприимства и развлечений для удобства граждан при планировании поездок и являются одним из условий улучшения качества туристских услуг.

Для ряда стран туризм также высокобюджетная составляющая, так как это мощнейшая индустрия, которая формирует порядка 10% мирового валового продукта, это сфера крупнейших инвестиций, которая предоставляет занятость миллионам людей разных профессий и квалификаций.

Стремительная цифровизация туристических продуктов на мировом и отечественном рынках, системы управления индустрией туризма, процессом, усугубленным вызовами новейшей истории, накладываемыми мировой пандемией, свидетельствует о том, что в ближайшее время произойдут глобальные изменения в отрасли. Выявление и исследование основных направлений происходящих трансформаций, сравнительный анализ предлагаемых цифровых решений в странах и регионах, технологий их реализации актуальны как для науки, так и для бизнес-сообщества как никогда прежде. Поскольку дальнейшее развитие туризма неразрывно связано с цифровыми трансформациями и информационными технологиями, назрела необходимость обратиться к его общей терминологии, трансформации и возможных моделей развития.

1. МЕТОДЫ МЕЖДИСЦИПЛИНАРНЫХ НАУЧНЫХ ИССЛЕДОВАНИЙ

Междисциплинарность в науке – одна из объективных закономерностей её современного развития. На практике результат любого исследования зависит от массива эмпирических данных; инструментария; правильно поставленных целей исследования. Междисциплинарный подход позволяет существенно расширить все компоненты, поскольку достижения одной дисциплины, включая тезаурус, накопленные методы, данные и результаты, могут быть использованы в качестве исходных данных или аппарата другой дисциплины. В современных научных ис-

следованиях актуализируется задача анализа перспективных междисциплинарных научных направлений, что позволяет прогнозировать результаты исследований в этих областях знаний и в различных сферах общественной жизни. Развитие науковедческих и научных дисциплин отстает от темпов роста терминологической и категориальной базы междисциплинарных научных областей, которая бесконтрольно формируется научными школами, группами и отдельными исследователями. Неоднозначность терминологии и неструктурированность значительной части информации, даже при свободном доступе, делает невозможным быстрый мониторинг новых тенденций.

Роль современных междисциплинарных научных областей и применимость традиционных научных методов в междисциплинарных исследованиях широко освещается в научных публикациях. Считается, что междисциплинарные области (фактически контексты) определяются посредством тематического поиска и описываются количественно и качественно с помощью различных видов измерений и процедур. Так, например, Okamura K., сосредотачиваясь на кластерах высоко цитируемых работ, широко известных как исследовательские направления (RF), предположил, что междисциплинарность статистически значима и положительно связана с исследовательским воздействием [2]. C. Carusi и G. Bianchi применяют количественную оценку междисциплинарности журналов, анализируя взаимосвязь между учеными и журналами, где они публикуются [3]. J. Raimbault предложил методологию измерения междисциплинарности, которая объединяет анализ цитированности и семантический анализ, определяющий качество связей между эндогенными дисциплинами [4]. Ученые G. Abramo и др. комплексно используют два библиометрических подхода к измерению междисциплинарных исследований: анализ дисциплинарного разнообразия в списке литературы; дисциплинарное разнообразие авторов [5]. C. Piciocchi и L. Martinelli исследуют категориально-концептуальные методы и подходы Digital Humanities для различных научных областей [6]. Методология исследовательских групп J. L. Jimenez-Marquez и др. и M. Pejic-Bacha и др. [7, 8] состоит из извлечения данных, аналитического подхода (описательный анализ и анализ текста) и анализа результатов применения машинного обучения. G. Paré и др. предлагают собственную типологию, фо-

кусируя исследования на подготовке и анализе обзоров [9]. Финские исследователи J. Hamari и J. Koivisto в своем исследовании практикуют интеллектуальный поиск и анализ научных текстов, предпочитая экспертную (ручную) обработку данных [10]. Исследователи Koivisto J. и Hamari J. предлагают использование системного подхода для изучения общественных наук, который, являясь аналитической основой, интегрирует результаты тематических исследований с целью расширения теоретической базы и их эмпирического понимания [10]. Стратегия поиска индонезийских авторов B. Purwandari и др., описанная в их статье [11], состоит в «сортировке цифровых библиотек, определении ключевых слов, использовании существующих инструментов в цифровых библиотеках для облегчения поиска и проведении первичных исследований».

Исследования в области цифрового, электронного, умного туризма проводятся уже более 10 лет и не теряют актуальности и востребованности. Более того, большинство аналитиков считает проводимые исследования недостаточными, отражающими только несколько наиболее заметных аспектов цифровых трансформаций и оставляющими в тени многочисленную тематику. Приступая к изучению данного междисциплинарного направления, интересно рассмотреть методологическую основу исследований в сфере туризма с целью уточнения собственных убеждений и наработок, а также согласования используемых науковедческих методов и практик поиска, извлечения и изучения контекстного знания. Обзор методов исследования, представленный ниже, проведен на подборке статей в Science Direct как результат комплексного запроса, включающего основные термин-концепты направления: электронный туризм, умный туризм и цифровой туризм (eTourism, smart tourism, digital tourism). Использование синтетического метода поиска и экспликации контекстов позволило сформировать подборку материалов, определить основные используемые методологии исследования в области цифрового туризма и базовые термин-концепты.

Обобщение аналитических методов, применяемых в исследованиях цифрового и умного туризма как междисциплинарных направлений, представлено в работе Jingjing Li и др. [12]. Так, в этой статье отмечено, что для извлечения и использования полезной информации, скрытой в онлайн-текстовых данных, в исследо-

ваниях в области туризма широко применяются разнообразные методы интеллектуального анализа текста, которые состоят из сбора и анализа данных. Собранные с помощью веб-сканера текстовые онлайн данные анализируются с целью извлечения полезных знаний (контекстных знаний) в два этапа: предварительная обработка данных и обнаружение шаблонов. Предварительная обработка включает операции очистки данных, токенизации, переноса по словам и маркировки частей речи. Результаты веб-поиска используются для прогнозирования в сфере туризма. Для этого предпринимаются два основных шага: выбор ключевых слов и введение предикторов, что позволяет в дальнейшем построить предсказательную модель. Выбор ключевых слов (терминологического ядра научного направления) является основным процессом в исследовании туризма с использованием данных веб-поиска, и результаты в значительной степени зависят от методов отбора. В исследованиях туризма широко используются три вида методов выбора ключевых слов: эмпирический (или экспериментальный), территориальный и технологический. Эмпирический подход определяет ключевые слова просто на основе знаний и опыта исследователей. Территориальный выбор ключевых слов является расширением эмпирического: сначала используется эмпирический метод для определения терминологического ядра, а затем добавляются связанные термин-концепты, относящиеся как к базовым, так и к рекомендованным (с использованием функции рекомендаций поисковых систем). Технологический метод отбирает ключевые слова из большой области выбора, основываясь на прогнозирующей способности, с точки зрения корреляции с прогнозирующими переменными.

Исследование Julio Navío-Marco и др. [13], по словам авторов, может быть классифицировано как нарративный обзор, направленный на анализ литературы и проведение критической оценки ее качества. Обзор проводился в традиционной манере, т. е. концептуальным и хронологическим образом. С методологической точки зрения использовались онлайновые базы научных публикаций (Web of Science, SciencieDirect) и различные комбинации ключевых слов, связанных с электронным туризмом. Так как в ряде предшествующих публикаций основным ограничением исследований было объявлено преимущественное использование

научных статей по туризму (что не соответствует междисциплинарности рассматриваемой тематики), авторы включили в обзор журналы, связанные с ИКТ. В исследовании была взята на вооружение следующая процедура: определены цель и объем (период, тематика обзора); определена процедура отбора материалов для обзора; произведена настройка процесса выбора источника (ручной поиск и идентификация релевантности с использованием различных комбинаций ключевых слов, включение в обзор статей JCR Q1 со значимым количеством цитирований); сформирована подборка наиболее релевантных материалов для анализа; произведен повторный отбор и обзор источников; выделен релевантный контент с использованием множественного кодирования (multiple coders). Авторы утверждают, что невозможно было подойти к обзору такого многоаспектного направления как «умный туризм», не ограничивая каким-либо образом объем материалов, учитывая наличие большое количество тем и подходов, которые отражают исследования связей между ИКТ и туризмом в последние годы.

В работе Jing Li и др. за методику исследования был взят «комплексный анализ критического медиа-дискурса (CMDA) путем интеграции анализа медиа-дискурса Carvalho (MDA, 2008) с критическим дискурс-анализом Fairclough (CDA, 1995). Подход CDA обеспечивает каркас скелета; в то время как схема анализа дискурса в СМИ предлагает аналитические компоненты для каждого измерения CDA» [14]. Набор данных включал новости, отраслевые отчеты и обзоры, журнальные статьи, обзоры экспертов, туристическую рекламу, передовые статьи и путевые заметки, собранные с помощью поисковой системы Google. CDA включает в себя три различных типа анализа: текстовый анализ, анализ обработки и социальный анализ, которые являются одновременными, но взаимозависимыми процессами. CDA подходит для исследования социальных и культурных изменений, поскольку облегчает интеграцию дискурс-анализа и анализа макро-контекстов. К текстовому описанию была применена схема текстового анализа Carvalho. Для интерпретации процессов использовался контекстный анализ. Исследование проводилось с использованием программного обеспечения Leximancer для выявления основных концепций и доминирующих тем в дискурсах СМИ с последующим ручным кодированием для проведения CMDA. Сбор данных начался с выбора ключевых слов. В качестве базовых термин-концептов, ключевых слов для

сбора данных медийных дискурсов были выбраны сочетания, которые наиболее часто появлялись в научных исследованиях. Определение базовых термин-концептов и их взаимосвязей является неотъемлемой частью дискурсивного анализа, а словарь, используемый для представления определенного явления, – существенным компонентом для экспликации значений. Пристальное внимание было уделено формированию тезауруса, включающего термин-концепты, встречающиеся в заголовках и первых абзацах статей в СМИ, что указывает на предпочтение в чтении. В результате в работе было выявлено 94 термин-концепта, которые были уточнены по 12 семантическим темам и отразили смысловое разнообразие контекстов в обрабатываемых материалах.

Цель систематического обзора литературы, представленного в работе Sanaz Shaee и др. [15], состояла в изучении различных аспектов интеллектуальных туристических направлений. Систематический обзор является явным и всеобъемлющим методом выявления, синтеза и оценки, а также объединения результатов существующих исследований по конкретной теме, представляющей интерес широкому кругу исследователей. В обзоре использовался метод обоснованной теории, нацеленный на исследование конкретного явления посредством индуктивного процесса, который позволяет генерировать теоретическое понимание явления. Подход полезен для проведения всестороннего теоретического анализа, связанного с темой, и заключается в объяснении целевого феномена в соответствии с понятиями, категориями и отношениями между ними. Выполнение анализа данных начинается с открытого кодирования (определение категорий, предложений и измерений), продолжается осевым кодированием (изучение стратегий, условий и последствий) и заканчивается выборочным кодированием (генерация теории). Первичные ссылки были выбраны на основе обзора заголовков, ключевых слов и рефератов. Запросы были сформированы с использование следующих ключевых слов: умный туризм, умные туристические направления, умный город, умный устойчивый город, информационно-коммуникационные технологии и устойчивый туризм. Затем заголовки, рефераты и ключевые слова статей были пересмотрены, чтобы найти другие термины и ключевые слова, используемые в исследовательской литературе, и разработать свою собственную подборку ключевых слов. Для получения лучших результатов поиска были использованы булевы

операторы: логическое «И» для объединения основных терминов и «ИЛИ» для включения синонимов.

Таким образом, в исследовательской среде сложились определённые подходы и методы, направленные на изучение контекстного знания на основе использования ИКТ. Эффективность их доказана анализом публикаций, отражающих их применения в различных междисциплинарных исследованиях.

2. МЕТОДОЛОГИЯ МЕЖДИСЦИПЛИНАРНЫХ НАУЧНЫХ ИССЛЕДОВАНИЙ

Для современных научных исследований характерен путь выявления контекстных знаний путем применения методов, подходов, технологий и инструментов цифровых гуманитарных наук, а также интеллектуального анализа текста. Контекстные знания обычно понимаются как способность правильно «читать» контекст, извлекать и интерпретировать профессионально значимую информацию посредством контекстного поиска. Контекстный поиск – это метод последовательного поиска текстовых фрагментов, относящихся к запросу пользователя. В этом случае контекст является частью текста, словесной средой выбранного элемента текста (термин-концепта) для анализа. Таким образом, понятие контекста интерпретируется нами как самостоятельная концептуальная единица категориального аппарата, которая может быть использована в качестве основы для классификации научных текстов, визуализации иерархических и ассоциативных отношений между терминами. Под термин-концептом подразумеваются одиночный термин или коллокация, несущие смысловую нагрузку определенного контекста.

Методология исследования основана на применении информационно-коммуникационных технологий в рамках разработанного авторами подхода (названного «синтетическим методом») к исследованиям развития различных предметных областей и практик человеческой деятельности через изучение формирования и развития их понятийно-терминологического аппарата. В ходе разработки и применения подхода в исследованиях проанализированы теоретические и практические результаты применения научно-исследовательских методов для извлечения, обработки и анализа контекстуальных знаний [16, 17]. Подход направлен на извлечение контекстных знаний из неструктурированных или полуструктурированных информационных ресурсов и позволяет посредством экспликации и

картирования формировать коллекции фрагментов, релевантных тематике (тематических контекстов). Подход комплексно обеспечивает применение методов поиска, извлечения, уточнения, экспликации, анализа и представления контекстного знания, построение трендов. Комплексность синтетического метода заключается в последовательном применении ИКТ на всех этапах исследования:

– отбор цифровых ресурсов, содержащих текстовые массивы, отражающие научный и общественно-политический дискурсы;

– использование аналитических инструментов отобранных ресурсов для экспликации контекстного знания и формирования контекстов вида «тематическая подборка», содержащих документы, релевантные исследуемой предметной области; на этом этапе также происходят экспертная оценка документов и на её основе качественный отбор максимально релевантных из них;

– применение аналитических инструментов цифровых ресурсов для получения статистических отчётов и их анализа;

– использование информационных систем и программного обеспечения для полнотекстового анализа сформированных тематических коллекций;

– применение программного обеспечения для обработки и интерпретации полученных данных контекстного анализа;

– построение трендов развития и формирование исследуемой предметной области.

Применение синтетического метода не зависит от выбора конкретных информационных систем и программного обеспечения, что обеспечивает гибкость в его использовании и доступность его применения в зависимости от возможностей исследовательских коллективов. Предлагаемый и используемый в исследовании подход согласуется с методами, принятыми в подобного рода исследованиях. Однако его особенность состоит как в учёте и интеграции различных применяемых методов, так и в собственных методах, к которым можно отнести, например:

– отказ от изучения тематической выборки высокоцитируемых научных журналов с высоким импакт-фактором в пользу рассмотрения более широкого круга публикаций из тематически различных изданий, что позволяет эксплициро-

вать большее число релевантных терминов, а также учитывать различные тенденции в развитии междисциплинарных направлений научных исследований, а не только самые распространённые;

– синтез различных методик, интегральный охват инструментов исследования и варьирование последовательности применения технологий поиска, отбора, экспликации и анализа контекстного знания в зависимости от начальных условий и особенностей конкретного исследования.

3. «ЦИФРОВОЙ ТУРИЗМ»: ТЕНДЕНЦИИ РАЗВИТИЯ

Зарубежные туристические компании активно и результативно ведут цифровизацию своей деятельности, получая значительные доходы.

Анализируя новостные ленты можно выделить четыре мировых тренда в туризме: платформы (platforms), экономика совместного использования (sharing economy), революцию впечатлений (experience revolution) и технологические гаджеты (technological gadgets).

Страны Азии, в экономику которых туризм вносит существенный вклад, приступили к цифровой трансформации на правительственном уровне. Например, Шри Ланка организовала масштабное цифровое промо страны как привлекательного туристического направления, а в Индонезии, в качестве ответа на международную экспансию «Airbnb», запущена собственная сеть бронирования гестхаусов и вилл «Indonesia Travel Exchange». Вслед за Японией, Китай и Южная Корея вкладывают значительные средства в развертывание технологической инфраструктуры, поддерживающей умный туризм. Речь идет прежде всего о создании насыщенного справочно-информационного поля, обеспечивающего потребности туриста на английском или другом доступном ему языке в режиме реального времени, там и тогда, где соответствующие потребности возникают.

В Европе «умные» дестинации нередко вырастают из ранее реализованных проектов Умного Города. Туристические приложения для мобильных телефонов возникают на основе существующих баз данных, подаваемых в новом ключе и для другой целевой аудитории.

В РФ туристская индустрия, ее развитие как на уровне государства в целом, так и для ее субъектов, муниципальных образований и общества, играют значи-

мую роль. Данные Ростуризма свидетельствуют, что из 4377 компаний, включенных в Единый федеральный реестр туроператоров, более 2,5 тыс. организаций осуществляют деятельность в сфере внутреннего туризма. Так как для страны важным является развитие въездного туризма, то ключевыми критериями при принятии решений о цифровизации данной сферы и развитии инфраструктуры должна быть ориентация на предпочтения потенциальных и фактических туристов. Среди важнейших цифровых решений можно выделить:

– создание туристского маркетплейса и централизацию усилий по продвижению туристского продукта Российской Федерации;

– внедрение и развитие мультиязычных сервисов помощи туристам, включая информационные сервисы, сервисы навигации и самообслуживания, с целью повышения доступности, качества и привлекательности туристских услуг, роста эффективности использования туристских ресурсов;

– разработку и реализацию электронной туристской карты гостя и аналогичного мобильного приложения в городах и субъектах Российской Федерации (аналог международных карт и приложений для мобильных устройств, позволяющих туристу перемещаться общественными видами транспорта, узнавать о культурных мероприятиях и событиях, пользоваться скидками при посещении объектов туристского показа, а также предоставляющих другие льготы);

– предоставление прозрачной электронной системы оценки качества предлагаемых туристских услуг, создание рейтинга туристских услуг и объектов по туристским территориям Российской Федерации;

– обеспечение возможности ознакомления с культурными и природными достопримечательностями, экспозициями музеев, туристскими маршрутами в онлайн-режиме с использованием технологий визуализации, виртуальных экскурсий, технологий дополненной реальности и др.;

– создание и развитие сервисов дополненной реальности для навигации по городам и объектам показа (музеям, выставочным центрам, художественным галереям и др.) для повышения привлекательности туристских объектов и эффективности использования туристских ресурсов;

– развитие системы открытых данных в сфере туризма для повышения прозрачности работы организаций и системы управления отраслью, создания условий для развития новых видов туристских услуг;

– внедрение и развитие технологий больших данных и искусственного интеллекта для сбора и анализа этих данных, а также развитие системы продвижения туристских услуг, формирование наиболее актуальных для туриста предложений с учетом его пожеланий, погодных условий, дорожной ситуации и др.;

– развитие сервисов онлайн-построения туристского маршрута с возможностью покупки билетов и бронирования гостиниц;

– создание электронной площадки для вовлечения самозанятых лиц в туристскую деятельность (гиды, инструкторы, экскурсоводы);

– разработку мультимедийных приложений для объектов показа, сервисов аудио- и видеогидов с возможностью интеграции с GPS-навигацией, использованием QR-кодов для формирования запросов.

Как видно из информации, представленной выше, задач и функций у современного туризма много. Из этого вытекают разные понимание и трактовка вообще понятия «цифровой туризм». В связи с этим возникает вопрос, а есть ли определенный интерес в научных кругах к данному направлению и каков он.

В целях анализа публикаций исследователей по тематике цифрового, электронного туризма были задействованы базы российской научной электронной библиотеки eLibrary и информационно-поисковой системы Google Scholar. Временной диапазон был выбран с 2010 по 2019 годы. При поиске в состав русскоязычной терминологической базы были заложены словосочетания «цифровой туризм», «digital tourism», «интеллектуальный туризм», «умный туризм», «smart tourism», «электронный туризм», «e-tourism». Выбор был обусловлен тем, что авторы при формировании статей пользуются как русскоязычным, так и англоязычным вариантами, что в свою очередь является требованием многих журналов.

Проведенный авторами анализ динамики русскоязычной терминологической базы цифрового туризма, полученных из НЭБ и Google Scholar, показывает:

– стремительный рост (за десятилетний период) активности научного интереса к тематике цифрового туризма, например, в 2019 году активность возросла почти в 30 раз (по данным eLibrary) и в 10 раз (по данным Google Scholar);

– тенденции и вектор развития по данным eLibrary и Google Scholar практически совпадают и имеют одинаковый характер (за исключением 2019 года, возможно, это связано с тем, что на период проведения исследования за 2019 год в российской научной электронной библиотеке пока еще не сформирована вся база публикаций);

– в общей тенденции выделяются два кластера (блока). Первый блок «затишье» – это период с 2010 по 2015 годы (количество публикаций незначительное и варьируется в пределах 2–12, особых изменений не наблюдается) и период «повышенной активности» – период после 2015 года по настоящее время (количество публикаций стремительно растет от года к году и варьируется в пределах 27–90). Причем относительно 2015 года активность по данным eLibrary увеличилась в 11 раз, а по данным Google Scholar – практически в 5 раз.

Таким образом, на основе полученных данных за десятилетний период, иллюстрирующих публикационную активность, можно утверждать о возрастающем научном интересе в общем плане – к тематике цифровизации, в частности – к тематике цифрового, электронного туризма, что ещё раз подтверждает актуальность исследования формирования терминологической базы «цифрового туризма».

3. ТЕРМИНОЛОГИЧЕСКОЕ ЯДРО «ЦИФРОВОГО ТУРИЗМА»

Одно из основных (классических) понятий туризма заложено в Федеральном законе «Об основах туристской деятельности в Российской Федерации», где данное понятие трактуется как «… временные выезды (путешествия) граждан Российской Федерации, иностранных граждан и лиц без гражданства (далее – лица) с постоянного места жительства в лечебно-оздоровительных, рекреационных, познавательных, физкультурно-спортивных, профессионально-деловых и иных целях без занятия деятельностью, связанной с получением дохода от источников в стране (месте) временного пребывания» [18]. Основы туристской деятельности, заложенные в этом определении, характеризуются широтой и масштабностью его элементов, охватывающих деятельность многих смежных сфер.

Как отмечалось выше, современный уровень цифровизации общества формирует новые формы коммуникационных взаимодействий и взаимоотношений между производителями и потребителями, в том числе в области туристических

услуг [19]. Производители туристских услуг вынуждены внедрять современные цифровые технологии, тем самым формируя новое направление – «цифровой туризм».

На сегодняшний день устоявшегося классического понятия «цифрового туризма» нет, в научных и публицистических вариантах интерпретация этого термина широка и разнообразна. Но в научном и медиа дискурсе предлагаются некоторые подходы и видение этого феномена через призму цифровизации. Так, в стратегии развития туризма до 2035 года можно выделить три группы понятий, непосредственно ассоциируемых с цифровизацией. Это цифровые технологии, цифровые решения и цифровые сервисы.

Цифровые технологии – интеграции информационных систем, данных, социальных платформ; мобильные; интернет; геймификации; визуализации виртуальных экскурсий; дополненной реальности и др.; больших данных; искусственного интеллекта; мультимедийные и т. п.

Цифровые решения – электронный документооборот; оцифровка; онлайн-платформа экосистемы туризма; электронная туристическая карта гостя; электронная система оценки качества услуг; электронные системы открытых данных; мультимедийные приложения; электронные площадки для самозанятых лиц; GPS-навигация; QR-коды.

Цифровые сервисы – электронные сервисы; мультиязычные информационные сервисы; сервисы навигации и самообслуживания; сервисы дополненной реальности; сервисы онлайн-построения туристического маршрута; сервисы покупки билета и бронирования гостиниц; сервисы аудио- и видеогидов.

Цифровая форма туризма в документе не определена явным образом, но анализ как зарубежной, так и отечественной литературы позволил выделить ряд базовых термин-концептов, определяющих процессы цифровизации в туризме: цифровой туризм, интеллектуальный туризм, электронный туризм и ряд других.

Проведенный выше анализ позволил предположить вариант терминологического ядра исследования. Терминологическое ядро исследуемого научного направления было сформировано с помощью аналитического аппарата Научной электронной библиотеки (НЭБ, http://elibrary.ru). Анализ основывался на стати-

стических данных о распределении ключевых слов в результатах запросов: «цифровой туризм» or «digital tourism» or «интеллектуальный туризм» or «умный туризм» or «smart tourism» or «электронный туризм» or «e-tourism». В данном случае в поиске использовались также англоязычные аналоги терминов. Это связано с тем, что они также встречаются в публикациях, а системой они рассматриваются как разные ключевые слова. В итоге первый уровень терминологического поля содержит – туризм и его вариант tourism; smart-туризм и smart-tourism; электронный туризм, e-Туризм и e-Tourism; цифровой туризм и digital tourism; цифровой номадизм и digital nomadism; цифровая экономика и digital economy; цифровые технологии и digital technologies; цифровая трансформация и digital transformation; цифровизация туризма; цифровая экосистема туризма. Учитывая, что в публикациях последних лет в ключевых словах добавилось большое количество терминов, таких как информационно-коммуникационные технологии, интернет, цифровые технологии, цифровые коммуникации, цифровизация, цифровая экономика, можно предположить, что туризм также превращается в цифровой.

4. БАЗОВЫЕ ТЕРМИН-КОНЦЕПТЫ ТЕРМИНОЛОГИЧЕСКОГО ЯДРА

На следующем этапе авторами был произведен анализ выявленных термин-концептов для целей включения их в терминологическое ядро. Качественный анализ был построен по отобранным экспертным образом научным публикациям и публикациям из СМИ с высокой степенью релевантности рассматриваемой предметной области. Отбор публикаций производился по трем текстовым массивам.

Первый массив содержал научные публикации, отобранные из российской Научной электронной библиотеки (НЭБ, http://elibrary.ru) и информационно-поисковой системы Google Scholar (https://scholar.google.ru).

Во второй вошли англоязычные публикации из информационных ресурсов ScienceDirect, Web of Science и Scopus.

Третий массив составили публикации из электронных архивов российских федеральных и региональных газет и журналов, а также интернет-изданий, представленных в информационной системе Интегрум (https://integrum.ru).

Исследования показали, что у авторов достаточно разный подход и к названию, и к трактовке, и к характеристике. Например, ряд авторов оперирует таким понятием, как «электронный туризм» (E-Tourism) и характеризуют его как «… не только электронная дистрибуция туруслуг, но и электронные экскурсии, которые также называют виртуальными» [20], отмечая, что появление этого термина связано «… с трансформацией термина «е-бизнес», который представляет собой применение широкого спектра возможностей ИКТ для организации полного цикла бизнес деятельности (е-коммерция, е-маркетинг, е-финансы, е-производство, е-стратегия, е-менеджмент)» [21], и он является «… частью электронной коммерции и объединяет быстроразвивающиеся сферы, такие как телекоммуникации и информационные технологии, в индустрию гостеприимства и управления» [22]. Как разновидность электронного туризма может выступать «мобильный туризм» (m-tourism) *«…* использующий мобильные технологии в виде приложений для мобильных телефонов (iPhone, iPad, Windows phone, Android) и позволяющий пользователям по телефону бронировать авиарейсы, отели, автомобили, находясь в любом месте» [20].

Стремительный рост цифровой грамотности и цифровых компетенций населения, развитие и широкомасштабное внедрение передовых технологий, таких как интернета вещей, больших данных, Wi-Fi, нейросетевых технологий и технологий 5G, позволяют туристу стать полноценным участником туриндустрии. Все это привело к новому понятию, так называемому «умному туризму» (smart-tourism). В современных публикациях отечественных и зарубежных авторов отмечается, что «умный туризм – набирающая силу тенденция, благодаря которой и местные жители, и туристы получают возможность взаимодействия с более удобной, безопасной, интересной средой обитания» [23], и «умный туризм – это модель объединенного развития туриндустрии и инновационной технологии S&R, что является не только будущей тенденцией развития туриндустрии, но и ключом к трансформации и модернизации современной сферы обслуживания» [24].

Существует трактовка «умного туризма» как туризм, «… при котором всесторонняя максимизация экологических, культурных, общественных и экономических ценностей может поощряться в качестве достижения устойчивого развития

сферы туризма с помощью таких информационных технологий, как интернет вещей, «облачные» вычисления, ГИС, виртуальная реальность и мобильный интернет» [24].

В сегодняшней практике широко используется слово «smart». Его применяют для описания технологических, социальных, экономических систем, активно внедряющих большие и открытые данные, интернет-технологии, всевозможные датчики, новые способы коммуникаций и обмена информацией. По аналогии с этим применение компьютера, ноутбука или смартфона для подготовки путешествия или во время оного формирует Smart-Tourism. По большому счету, это умение получить туристическую услугу через интернет в любой точке мира, на любом языке.

Как показало исследование, ряд отечественных авторов в своих публикациях трактуют smart tourism и как «интеллектуальный туризм», подразумевая под этим «… туризм, поддерживаемый на уровне туристического региона интегрированными усилиями по поиску инновационных способов накопления и агрегирования или использования данных, извлеченных из инфраструктуры, социальных связей, государственных или организационных источников» [25] и «… туризм, в котором постоянное и систематическое использование умных элементов приводит к созданию дополнительной ценности путешествия для туриста» [26].

Нынешнее мировое производство характеризуется сменой технологического уклада, а именно, Четвертой индустриальной революцией. 2011 год – год, когда в научный оборот был введен термин «Индустрия 4.0». В общем плане понятие Четвертой индустриальной революции (Индустрия 4.0) трактуется как переход на полностью автоматизированное цифровое производство, управляемое интеллектуальными системами в режиме реального времени в постоянном взаимодействии с внешней средой, выходящее за границы одного предприятия, с перспективой объединения в глобальную промышленную сеть Вещей и услуг [27]. Четвертая индустриальная революция изменяет способ ведения бизнеса не только в промышленности, но и во всех секторах. Не обошла она и туристическую область, выделив в ней «Туризм 4.0».

Таким образом, в научных трудах появилась следующая трактовка «Туризма 4.0» – «это наименование современной концепции обработки больших

данных, собранных в результате исследования различных туристских дестинаций, для создания персонализированного информационного пространства туристских ресурсов» [28]. «Туризм 4.0», основываясь на механизмах Индустрии 4.0, способствует «… развитию туристских дестинаций региона и позволяет разработать эффективную туристическую политику по средствам процессов цифровизации и автоматизации» [28] и помогает населению совершать путешествия, делая эти поездки увлекательными, эффективными, безопасными и персонализированными.

На основе полученных данных по русскоязычному дискурсу можно сделать вывод, что цифровизация в туризме неизбежна и позволяет сделать туристический бизнес более гибким, адаптированным к реалиям современности и конкурентоспособным в развивающемся «цифровом мире».

На следующем шаге исследования из полнотекстовой базы научной информации Science Direct были выявлены термин-концепты компаньоны ключевых слов «электронный туризм», «умный туризм» и «цифровой туризм» (eTourism, smart tourism, digital tourism), используемых в запросах: intelligent tourism, digital free tourism (DFT), sustainable tourism. Все вместе они составляют одну семантическую группу: цифровые формы туризма. Поиск термин-концептов терминологического ядра осуществлялся по библиографическому описанию (название публикации, ключевые слова, аннотация), полному тексту публикации и списку источников.

На следующем шаге исследования был проведен контент-анализ отобранных текстов, экспликация контекстного знания и была уточнена смысловая нагрузка для каждого из термин-концептов, определяющих цифровые формы туризма. Развитие терминологии, сущности и понятия цифрового туризма в ретроспективе можно представить следующим образом:

Sustainable tourism – «удовлетворение потребности нынешних туристов и принимающих регионов при одновременной защите и расширении возможностей на будущее; … управление всеми ресурсами таким образом, чтобы экономические, социальные и эстетические потребности могли быть удовлетворены при сохранении культурной целостности, основных экологических процессов, биологического разнообразия и систем жизнеобеспечения» (UNWTO [29], 1993 г.); «кон-

троль и локальное планирование туристических процессов; … достижение максимальной эффективности в потреблении ресурсов и минимизации воздействия туристической деятельности на окружающую среду» (UNWTO [30], 2013 г.); «природная и культурная устойчивость, экономические показатели и конкурентоспособность, занятость и человеческий капитал, сокращение бедности, управление социальной инклюзией; … концепции устойчивости (sustainable) и разумности (smart) имеют много общих элементов. На концептуальном уровне первое подразумевается во втором. То есть туризм нельзя считать умным, если он не устойчивый» (José Francisco Perles Ribes, Josep Ivars Baidal [31], 2018 г.);

Smart tourism – «целостный, долгосрочный и устойчивый подход к планированию, разработке, эксплуатации и маркетингу туристических продуктов и бизнеса. Умный туризм формируется двумя типами техник: 1) умный спрос и использование методов управления, которые способны управлять спросом и доступом; 2) умные маркетинговые методы, которые можно использовать для определения целевых сегментов клиентов для доставки соответствующих сообщений» (S.G. Phillips, [32]; 2000 г.); «использование мобильной цифровой связи для создания более интеллектуальных, значимых и устойчивых связей между туристами и пунктами назначения; форма гражданского участия, а не просто форма потребления» (J.G. Molz, [33], 2012 г.); «чистые, экологичные, этические и высококачественные сервисы, предлагаемые на всех уровнях цепочки услуг» (UNWTO [34], 2012 г.); «отдельная система поддержки туристов в контексте информационных сервисов и всеобъемлющих технологий (Y. Li et al. [35], 2017 г.); «связан с устройствами, генерирующими большие данные различной природы для управления туризмом» (Jingjing Li et al. [12], 2018 г.); «замена большей части человеческого труда в индустрии путешествий, туризма и гостеприимства цифровыми технологиями» (Julio Navío-Marco et al. [13], 2018 г.); «умный туризм похож на термин «интеллектуальный туризм», но в отличие от последнего содержание «умного» более обширно и требует большого объема данных; «разумность» делает больший упор на технологические результаты для людей, «интеллект» лежит в основе полезности знаний и информации; «разумность» — это сублимация интеллектуальных сил, предвосхищающих потребности», «вездесущий информационный сервис (Smart tourism is ubiquitous tour information service), предоставляемый отдельным

туристам, а не туристическим группам, получаемый туристами во время туристического процесса в любое время, в любом месте и на основе индивидуальных требований людей» (José Francisco Perles Ribes, Josep Ivars Baidal [31], 2018 г.); «туристические продукты, в которых используются технологические компоненты» (Inta Egger et al. [36], 2020 г.); «логическое эволюционное развитие традиционного туризма и электронного туризма, в котором заложена основа для инноваций, основанных на технологиях» (Aristea Kontogianni, Efthimios Alepis [5], 2020 г.); «ключевая концепция, включающая: сохранение конфиденциальности, осведомленность о контексте, культурное наследие, систему рекомендаций, социальные сети, интернет вещей, пользовательский опыт, системы реального времени, моделирование профилей пользователей, дополненную реальность и большие данные» (Aristea Kontogianni, Efthimios Alepis [5], 2020 г.);

intelligent tourism – «способность изменять свое состояние или действие в ответ на различные ситуации, различные требования и предыдущий опыт; сосредоточен на сфере технологий, технических возможностях, предлагает пользователям более удобные и эффективные услуги (включая материальные продукты)» (José Francisco Perles Ribes, Josep Ivars Baidal [31], 2018 г.);

digital tourism – «сближение между физическим и цифровым мирами, поддерживаемое датчиками, которые собирают данные, возникающие в результате взаимодействия между туристами и окружающей средой» (Julio Navío-Marco, et al. [13], 2018 г.);

digital free tourism (DFT) – «описывает туристические пространства, в которых интернет или мобильные сигналы либо отсутствуют, либо использование цифровых технологий контролируется»; «характеризуется отсутствием или серьезным ограничением доступа к информационным и коммуникационным технологиям (ИКТ)» (Jing Li et al. [14], 2018 г.);

eTourism – «исследования проявлений туризма посредством ИКТ в широком смысле» (Jingjing Li et al. [12], 2018 г.); «ИКТ в управлении туризмом» (Julio Navío-Marco et al. [13], 2018 г.); «один из результатов внедрения технологий в туристической индустрии; современные технологии уже не просто инструмент для электронного туризма, а используются во всех аспектах жизни и путешествий» (Sanaz Shaee et al. [15], 2019 г.).

В целом, как показал анализ, термин «digital tourism» не является самым распространённым в зарубежных исследованиях. В академической среде существует плюрализм, который, как видится, поддерживается разными подходами к исследованиям, пристрастиям авторов, а также отсутствием каких-либо стандартов на международном уровне, которые хоть как-то регулировали использование терминологии.

На следующем шаге с помощью информационно-поисковой системы Интегрум были получены результаты по запросам («цифровой туризм» или «электронный туризм») и («умный туризм» или «смарт туризм»). Полученные результаты говорят о достаточно слабом интересе официальных СМИ к развитию сферы туризма на основе использования информационно-коммуникационных технологий. Так, в период с 2010 по 2019 годы по первому запросу было выявлено всего 18, а по второму – 17 публикаций при явно хаотичном разбросе по годам.

Анализ самих публикаций позволяет сделать следующие выводы:

– цифровой туризм воспринимается в обществе как новое и ещё малоизученное явление без чёткого понимания сущности этого понятия;

– в образовательной сфере появляются новые направления подготовки специалистов и повышения квалификации госслужащих в сфере «цифрового туризма»;

– появляются и начинают реализовываться инициативные проекты (в основном на региональном уровне) в русле «цифрового туризма» и «умного туризма» (в большей степени), например, различные специализированные интернет-порталы, агрегаторы и поисковые системы;

– «цифровой туризм» и «умный туризм» рассматриваются в качестве механизмов привлечения людей в сферу реального туризма (реализация механизмов партисипативности);

– обсуждение развития «цифрового туризма» и «умного туризма» происходит на различных региональных и международных площадках (форумах и фестивалях).

В целом дискурс в СМИ показывает, что журналисты используют терминологию «цифрового туризма» скудно. При этом особо не утруждают себя серьёзно погрузиться в эту тематику. В основном в заголовке или аннотации одноразово

используется термин «цифровой туризм» или «умный туризм», а затем по тексту никаких терминов, отражающих тематику «цифрового туризма», не встречается.

ЗАКЛЮЧЕНИЕ

Проведенное авторское исследование по предлагаемой методике позволило: выявить тенденции и открыть новые возможности для исследований, особенно в тех областях, где до сих пор не достаточно академических исследований и анализа научной литературы по цифровым трансформациям туризма, связанных с применением современных цифровых технологий, таких как большие данные, искусственный интеллект, виртуальная реальность, умные технологии и др. Благодаря информационно-коммуникационным технологиям, повсеместной и всеобщей доступности, современные технологии уже не просто инструмент для электронного туризма, а возможность широкого использования во всех аспектах жизни и путешествий.

Актуальность результатов проведённого исследования заключается в том, что замена в самом ближайшем будущем большей части человеческого труда в индустрии туризма и гостеприимства на цифровые технологии будет иметь последствия, которые стоит изучать уже сегодня. При этом авторы предлагают собственные методы и подходы к этим исследованиям.

Благодарности

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект 18-011-00923-а.

СПИСОК ЛИТЕРАТУРЫ

1. Распоряжение Правительства РФ от 20.09.2019 № 2129-р. «Об утверждении Стратегии развития туризма в Российской Федерации на период до 2035 года».
URL:http://static.government.ru/media/files/FjJ74rYOaVA4yzPAshEu-lYxmWSpB4lrM.pdf.

2. *Okamura K.* Interdisciplinarity revisited: evidence for research impact and dynamism // Palgrave Communication. 2019. V. 5. 141.
https://doi.org/10.1057/s41599-019-0352-4

3. *Carusi C., Bianchi G.* A look at interdisciplinarity using bipartite scholar/

journal networks // Scientometrics. 2020. No. 122. P. 867–894.
https://doi.org/10.1007/s11192-019-03309-3

4. *Raimbault J.* Exploration of an interdisciplinary scientific landscape // Scientometrics. 2019. V. 119. No. 2. P. 617–641. https://doi.org/10.1007/s11192-019-03090-3

5. *Kontogianni A., Alepis E.* Smart tourism: State of the art and literature review for the last six years // Array. 2020. V. 6. 100020.
https://doi.org/10.1016/j.array.2020.100020.

6. *Piciocchi C., Martinelli L.* The change of definitions in a multidisciplinary landscape: the case of human embryo and preembryo identification // Croatian Medical Journal. 2016. V. 57. No. 5. P. 510–515.
https://doi.org/10.3325/cmj.2016.57.510

7. *Jimenez-Marquez J.L., Gonzalez-Carrasco I., Lopez-Cuadrado J.L., Ruiz-Mezcua B.* Towards a big data framework for analyzing social media contente // International Journal of Information Management. 2019. V. 44. P. 1–12.
https://doi.org/10.1016/j.ijinfomgt.2018.09.003.

8. *Pejic-Bacha M., Bertoncelb T., Meškob M., Krstićc Ž.* Text mining of industry 4 job advertisements // International Journal of Information Management. 2020. V. 50. P. 416–431. https://doi.org/10.1016/j.ijinfomgt.2019.07.014

9. *Paré G., Trudel M. C., Jaana M., Kitsiou S.* Synthesizing information systems knowledge: A typology of literature reviews // Information & Management. 2015. V. 52. No. 2. P. 183–199. https://doi.org/10.1016/j.im.2014.08.008

10. *Koivisto J., Hamari J.* The rise of motivational information systems: A review of gamification research // International Journal of Information Management. 2019. V. 45. P. 191–210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013

11. *Purwandari B., Sutoyo M.A.H., Mishbah M., Dzulfikar M.F.* Gamification in e-Government: A Systematic Literature Review // Fourth International Conference on Informatics and Computing (ICIC). Semarang, Indonesia, 2019. P. 1–5.
https://doi.org/10.1109/ICIC47613.2019.8985769

12. *Li J., Xu L., Tang L., Wang S., Li L.* Big data in tourism research: A literature review // Tourism Management. 2018. No. 68. 301e323.
https://doi.org/10.1016/j.tourman.2018.03.009

13. *Navío-Marco J., Ruiz-Gómez L.M., Sevilla-Sevilla C.* Progress in information technology and tourism management: 30 years on T and 20 years after the internet – Revisiting Buhalis & Law's landmark study about eTourism // Tourism Management. 2018. No. 69. P. 460–470. https://doi.org/10.1016/j.tourman.2018.06.002

14. *Li J., Pearce P.L., Low D.* Media representation of digital-free tourism: A critical discourse analysis // Tourism Management. 2018. No. 69. P. 317–329. https://doi.org/10.1016/j.tourman.2018.06.027

15. *Shaee S., Ghatari A.R., Hasanzadeh A., Jahanyan S.* Developing a model for sustainable smart tourism destinations: A systematic review // Tourism Management Perspectives. 2019. No. 31. P. 287–300. https://doi.org/10.1016/j.tmp.2019.06.002

16. *Кононова О.В., Ляпин С.Х., Прокудин Д.Е.* Исследование терминологической базы междисциплинарного научного направления «цифровая экономика» с использованием инструментов контекстного анализа // International Journal of Open Information Technologies. 2018. Т. 6, № 12. С. 57–66.

17. *Кононова О.В., Прокудин Д.Е., Смирнова П.В.* Технологии изучения контекстного знания при исследованиях основных направлений геймификации в городском развитии // Информационное общество: образование, наука, культура и технологии будущего. Выпуск 3 (Труды XXII Международной объединенной научной конференции «Интернет и современное общество», IMS-2019, Санкт-Петербург, 19–22 июня 2019 г. Сборник научных трудов). СПб: Университет ИТМО, 2019. С. 53–66. https://doi.org/10.17586/2587-8557-2019-3-53-66.

18. Федеральный закон от 24.11.1996 N 132-ФЗ (ред. от 03.07.2019, от 01.04.2020). «Об основах туристской деятельности в Российской Федерации». URL: http://www.consultant.ru/document/cons_doc_LAW_12462/bb9e97fad9d14ac66df4b6e67c453d1be3b77b4c/

19. *Черевичко Т.В., Темякова Т.В.* Цифровизация туризма: формы проявления // Изв. Сарат. ун-та. Нов. сер. Сер. Экономика. Управление. Право. 2019. Т. 19. Вып. 1. С. 59–64.

20. *Мошняга Е.В.* Основные тенденции развития туризма в современном мире // Вестник РМАТ. 2013. № 3(9). С. 20–33.

21. *Калмакова А.А.* Цифровые туристические экосистемы и их роль в маркетинге дестинаций // География и туризм: сб. науч. тр. / Перм. гос. нац. исслед.

ун-т. Пермь, 2015. Вып. 14. С. 57–62.

22. *Устинова М.В., Шевченко М.В.* Индустрия гостеприимства в эпоху цифровизации // Эпоха науки. 2019. № 20. С. 459–463.

23. *Молчанова В.А.* Тенденции инновационного развития туристских дестинаций: «умная дестинация» // Экономика и предпринимательство. 2017. № 9 (ч. 3). С. 715–720.

24. *Сяоцянь К., Шуцинь А.* Исследование развития «умного» туризма в провинции Цзянси в рамках концепции «Интернет+» // Экономические и социальные перемены в регионе: факты, тенденции, прогноз. 2016. № 4. С. 199–205.

25. *Смирнов А.В., Пономарев А.В., Левашова Т.В., Тесля Н.Н.* Поддержка принятия решений в туризме на основе человеко-машинного облака // Искусственный интеллект и принятие решений. 2017. № 2. С. 90–102.

26. *Кормягина Н.Н.* Smart-туризм как часть Smart-концепции // Маркетинг и логистика. 2017. №6 (14). С. 45–57.

27. Четвертая промышленная революция. Популярно о главном технологическом тренде XXI века // TAdviser. 17.10.2017.
URL: http://www.tadviser.ru/index.php/ Статья:Четвертая_промышленная_революция_(Industry_Индустрия_4.0).

28. *Щедрина Е.Ю., Моисеева А.Г., Гончаров А.Н., Хубулова В.В.* Цифровой туризм: как индустрия 4.0 повлияет на туристическую отрасль региона // Современная наука и инновации. 2019. №1. С. 251–256.

29. UNWTO (1993): Tourism the year 2000 and beyond qualitative aspects, Madrid, OMT. Madrid. URL: https://www.unwto.org.

30. UNWTO (2013): Sustainable Tourism for Development Guidebook. Enhancing capacities for Sustainable Tourism for development in developing countries, Madrid, UNWTO. URL: https://www.unwto.org.

31. *Perles Ribes J.F., Baidal J.I.* Smart sustainability: a new perspective in the sustainable tourism debate // Investigaciones Regionales – Journal of Regional Research. 2018. No. 42. P. 151–170.

32. *Phillips S.G.* The tourism industry association of Canada [EB/OL] // URL: http://www.slideshare.com. 2000-12-05.

33. *Molz J.G.* Travel connections: Tourism, technology and togetherness in a

mobile world. London: New York, Routledge, 2012.

34. UNWTO. Tourism resilience committee stresses need for “Smart Tourism” [EB/OL]. 2012-03-11. URL: http://www.slideshare.com.

35. *Li Yu., Hu C., Huang Ch., Duan L.* The concept of smart tourism in the context of tourism information services // Tourism Management. 2017. No. 58. P. 293–300. https://doi.org/10.1016/j.tourman.2016.03.014

36. *Egger I., Lei S.I., Wassler P.* Digital free tourism – An exploratory study of tourist motivations // Tourism Management. 2020. No. 79. 104098. https://doi.org/10.1016/j.tourman.2020.104098

RESEARCH OF THE CONTEXTS OF THE ECOSYSTEM OF "DIGITAL TOURISM"

O. V. Kononova[1, 4, [0000-0001-6293-7243]], D. E. Prokudin[1, 2, 4, [0000-0002-9464-8371]],

E. N. Tupikina[3,4, [0000-0001-9531-9900]]

[1] *ITMO University, St. Petersburg, Russia*

[2] *Saint-Petersburg State University, St. Petersburg, Russia*

[3] *Far Eastern Federal University, Vladivostok, Russia*

[4] *Center digital society research, Russia*

[1]kononolg@yandex.ru, [2]hogben.young@gmail.com, [3]etupikina@mail.ru

Abstract

Modern information and communication technologies, elements of digitalization are constantly and rapidly developing, which, in turn, has a direct impact on all spheres of human activity. In the light of recent events related to the collapse of the tourism business due to COVID-19, there is a great scientific interest in the service sector, namely in the field of "digital tourism". Digital tourism relies on the widespread adoption of new technologies such as social media and mobile technologies, smart devices and sensors to collect and use massive amounts of data to create new value propositions. In this regard, the authors set a goal – to present a review of the literature on "digital tourism" from the standpoint of scientific and media discourse. The authors

present a comprehensive scientific approach, including the sequential implementation of all stages of the review, from the definition of the terminological core of the interdisciplinary direction, the formation of search queries, cascade search, selection and content analysis of materials to the identification and explication of contexts. The sources of information for preparing the review were publications from academic databases: Web of Science, Science-Direct, Scopus, GoogleScholar, eLibrary, Cyberleninka, as well as materials and publications in Russian-language media – Integrum.

The results obtained by the authors will be useful for scientists in identifying promising areas of research in the field of "digital tourism", as well as deepen their knowledge of mechanisms for searching, collecting and analyzing data and integrated and analytical environments.

Keywords*: information and communication technology, digital transformations, digital tourism, eTourism, smart tourism.*

REFERENCES

1. Rasporjazhenie Pravitel'stva RF ot 20.09.2019 № 2129-r. «Ob utverzhdenii Strategii razvitija turizma v Rossijskoj Federacii na period do 2035 goda». URL: http://static.government.ru/media/files/FjJ74rYOaVA4yzPAshEulYxmWSpB4lrM.pdf

2. *Okamura K.* Interdisciplinarity revisited: evidence for research impact and dynamism // Palgrave Communication. 2019. V. 5. 141. https://doi.org/10.1057/s41599-019-0352-4

3. *Carusi C., Bianchi G.* A look at interdisciplinarity using bipartite scholar/ journal networks // Scientometrics. 2020. No. 122. P. 867–894. https://doi.org/10.1007/s11192-019-03309-3

4. *Raimbault J.* Exploration of an interdisciplinary scientific landscape // Scientometrics. 2019. V. 119. No. 2. P. 617–641. https://doi.org/10.1007/s11192-019-03090-3

5. *Kontogianni A., Alepis E.* Smart tourism: State of the art and literature review for the last six years // Array. 2020. V. 6. 100020. https://doi.org/10.1016/j.array.2020.100020.

6. *Piciocchi C., Martinelli L.* The change of definitions in a multidisciplinary landscape: the case of human embryo and preembryo identification // Croatian

Medical Journal. 2016. V. 57. No. 5. P. 510–515.
https://doi.org/10.3325/cmj.2016.57.510

7. *Jimenez-Marquez J.L., Gonzalez-Carrasco I., Lopez-Cuadrado J.L., Ruiz-Mezcua B.* Towards a big data framework for analyzing social media contente // International Journal of Information Management. 2019. V. 44. P. 1–12.
https://doi.org/10.1016/j.ijinfomgt.2018.09.003.

8. *Pejic-Bacha M., Bertoncelb T., Meškob M., Krstićc Ž.* Text mining of industry 4 job advertisements // International Journal of Information Management. 2020. V. 50. P. 416–431. https://doi.org/10.1016/j.ijinfomgt.2019.07.014

9. *Paré G., Trudel M. C., Jaana M., Kitsiou S.* Synthesizing information systems knowledge: A typology of literature reviews // Information & Management. 2015. V. 52. No. 2. P. 183–199. https://doi.org/10.1016/j.im.2014.08.008

10. *Koivisto J., Hamari J.* The rise of motivational information systems: A review of gamification research // International Journal of Information Management. 2019. V. 45. P. 191–210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013

11. *Purwandari B., Sutoyo M.A.H., Mishbah M., Dzulfikar M.F.* Gamification in e-Government: A Systematic Literature Review // Fourth International Conference on Informatics and Computing (ICIC). Semarang, Indonesia, 2019. P. 1–5.
https://doi.org/10.1109/ICIC47613.2019.8985769

12. *Li J., Xu L., Tang L., Wang S., Li L.* Big data in tourism research: A literature review // Tourism Management. 2018. No. 68. 301e323.
https://doi.org/10.1016/j.tourman.2018.03.009

13. *Navío-Marco J., Ruiz-Gómez L.M., Sevilla-Sevilla C.* Progress in information technology and tourism management: 30 years on T and 20 years after the internet – Revisiting Buhalis & Law's landmark study about eTourism // Tourism Management. 2018. No. 69. P. 460–470. https://doi.org/10.1016/j.tourman.2018.06.002

14. *Li J., Pearce P.L., Low D.* Media representation of digital-free tourism: A critical discourse analysis // Tourism Management. 2018. No. 69. P. 317–329.
https://doi.org/10.1016/j.tourman.2018.06.027

15. *Shaee S., Ghatari A.R., Hasanzadeh A., Jahanyan S.* Developing a model for sustainable smart tourism destinations: A systematic review // Tourism Management Perspectives. 2019. No. 31. P. 287–300. https://doi.org/10.1016/j.tmp.2019.06.002

16. *Kononova O.V., Lyapin S.Kh., Prokudin D.E.* Studying the Interdisciplinary Terminological Landscape of Digital Economy with the Use of Contextual Analysis Tools // International Journal of Open Information Technologies. 2018. V. 6, No. 12. P. 57–66.

17. *Kononova O.V., Prokudin D.E., Smirnova P.V.* Approach to Use of Network Scientific Environment for Studying the Interdisciplinary Terminological Landscape of Digital Economy // Information Society: Education, Science, Culture and Technology of Future. Issue 3. P. 53–66. https://doi.org/10.17586/2587-8557-2019-3-53-66.

18. Federal'nyj zakon ot 24.11.1996 N 132-FZ (red. ot 03.07.2019, ot 01.04.2020). "Ob osnovah turistskoj dejatel'nosti v Rossijskoj Federacii". URL: http://www.consultant.ru/document/cons_doc_LAW_12462/bb9e97fad9d14ac66df4b6e67c453d1be3b77b4c/

19. *Cherevichko T.V., Temjakova T.V.* Cifrovizacija turizma: formy projavlenija // Izv. Sarat. un-ta. Nov. ser. Ser. Jekonomika. Upravlenie. Pravo. 2019. T. 19. Vyp. 1. S. 59–64.

20. *Moshnjaga E.V.* Osnovnye tendencii razvitija turizma v sovremennom mire // Vestnik RMAT. 2013. № 3(9). S. 20–33.

21. *Kalmakova A.A.* Cifrovye turisticheskie jekosistemy i ih rol' v marketinge destinacij // Geografija i turizm: sb. nauch. tr. / Perm. gos. nac. issled. un-t. Perm', 2015. Vyp. 14. S. 57–62.

22. *Ustinova M.V., Shevchenko M.V.* Industrija gostepriimstva v jepohu cifrovizacii // Jepoha nauki. 2019. № 20. S. 459–463.

23. *Molchanova V.A.* Tendencii innovacionnogo razvitija turistskih destinacij: «umnaja destinacija» // Jekonomika i predprinimatel'stvo. 2017. № 9 (ch. 3). S. 715–720.

24. *Sjaocjan' K., Shucin' A.* Issledovanie razvitija «umnogo» turizma v provincii Czjansi v ramkah koncepcii «Internet+» // Jekonomicheskie i social'nye peremeny v regione: fakty, tendencii, prognoz. 2016. № 4. S. 199–205.

25. *Smirnov A.V., Ponomarev A.V., Levashova T.V., Teslja N.N.* Podderzhka prinjatija reshenij v turizme na osnove cheloveko-mashinnogo oblaka // Iskusstvennyj intellekt i prinjatie reshenij. 2017. № 2. S. 90–102.

26. *Kormjagina N.N.* Smart-turizm kak chast' Smart-koncepcii // Marketing i logistika. 2017. №6 (14). S. 45–57.

27. Chetvertaja promyshlennaja revoljucija. Populjarno o glavnom tehnologicheskom trende XXI veka // TAdviser. 17.10.2017. URL: http://www.tadviser.ru/index.php/Статья:Четвертая_промышленная_революция_(Industry_Индустрия_4.0).

28. *Shhedrina E.Ju., Moiseeva A.G., Goncharov A.N., Hubulova V.V.* Cifrovoj turizm: kak industrija 4.0 povlijaet na turisticheskuju otrasl' regiona // Sovremennaja nauka i innovacii. 2019. №1. S. 251–256.

29. UNWTO (1993): Tourism the year 2000 and beyond qualitative aspects, Madrid, OMT. Madrid. URL: https://www.unwto.org.

30. UNWTO (2013): Sustainable Tourism for Development Guidebook. Enhancing capacities for Sustainable Tourism for development in developing countries, Madrid, UNWTO. URL: https://www.unwto.org.

31. *Perles Ribes J.F., Baidal J.I.* Smart sustainability: a new perspective in the sustainable tourism debate // Investigaciones Regionales – Journal of Regional Research. 2018. No. 42. P. 151–170.

32. *Phillips S.G.* The tourism industry association of Canada [EB/OL] // URL: http://www.slideshare.com. 2000-12-05.

33. *Molz J.G.* Travel connections: Tourism, technology and togetherness in a mobile world. London: New York, Routledge. 2012.

34. UNWTO. Tourism resilience committee stresses need for “Smart Tourism” [EB/OL]. 2012-03-11. URL: http://www.slideshare.com.

35. *Li Yu., Hu C., Huang Ch., Duan L.* The concept of smart tourism in the context of tourism information services // Tourism Management. 2017. No. 58. P. 293–300. https://doi.org/10.1016/j.tourman.2016.03.014

36. *Egger I., Lei S.I., Wassler P.* Digital free tourism – An exploratory study of tourist motivations // Tourism Management. 2020. No. 79. 104098. https://doi.org/10.1016/j.tourman.2020.104098

СВЕДЕНИЯ ОБ АВТОРАХ

КОНОНОВА Ольга Витальевна – кандидат экономических наук, доцент Национального исследовательского университета ИТМО;

Olga KONONOVA – Associate professor, PhD (Economy), ITMO University, Saint-Petersburg, Russia;

E-mail: kononolg@yandex.ru

ПРОКУДИН Дмитрий Евгеньевич – доктор философских наук, доцент, доцент Санкт-Петербургского государственного университета, аналитик Национального исследовательского университета ИТМО;

Dmitry PROKUDIN – Associate professor, Dr. Science (Philosophy), Saint-Petersburg State University, ITMO University, Saint-Petersburg, Russia;

E-mail: hogben.young@gmail.com

ТУПИКИНА Елена Николаевна – кандидат экономических наук, доцент Дальневосточного федерального университета;

Elena TUPIKINA – Associate professor Far Eastern Federal University

E-mail: etupikina@mail.ru

Материал поступил в редакцию 26 ноября 2020 года

УДК 51-77

ОПРОВЕРЖЕНИЕ СЛУХА СРЕДСТВАМИ МАССОВОЙ ИНФОРМАЦИИ: МАТЕМАТИЧЕСКАЯ МОДЕЛЬ И ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ

А. П. Михайлов[1, [0000-0002-2730-1538]], **А. П. Петров**[2, [0000-0001-5244-8286]]

Институт прикладной математики им. М.В. Келдыша Российской академии наук, г. Москва

[1]apmikhailov@yandex.ru, [2]petrov.alexander.p@yandex.ru

Аннотация

Рассмотрен процесс, при котором в социуме распространяется недостоверный слух, которому противодействует вещание средств массовой информации. Недостоверность слуха в данном случае понимается так, что информация СМИ содержит опровержение и тем самым инокулирует индивидов, то есть делает их невосприимчивыми к слуху. В то же время индивиды, успевшие принять слух, перестают доверять средствам массовой информации и тем самым становятся недоступными для переубеждения. Для данного процесса предложена математическая модель в двух вариантах. Вариант с непрерывным временем позволяет выявить некоторые математические свойства модели. Вариант с дискретным временем более удобен для анализа реальных процессов, так как позволяет оценить параметры модели. Для оценки этих параметров использованы данные о рейтингах основных социально-политических программ российских телеканалов. Приведено несколько сценарных расчетов модели с этими параметрами. Основной вывод состоит в том, что если информация, распространяемая средствами массовой информации, не является вирусной, то есть не пересказывается зрителями своим соседям по социуму, то СМИ оказываются не в состоянии противостоять слухам.

Ключевые слова*: математическое моделирование, информационное противоборство, численный эксперимент, слухи.*

ВВЕДЕНИЕ

Одна из актуальных форм современного информационного противоборства – это борьба средств массовой информации против «сенсационного» быстро распространяющегося слуха. Следствием «сенсационности» является вирусность, что означает передачу слуха от одного носителя достаточно большому количеству его соседей по социуму. Средства массовой информации могут инокулировать индивида, т. е. сделать его невосприимчивым заражению слухом, предложив ему критические аргументы. Однако если индивид уже воспринял слух и стал его адептом, то, наоборот, он скептически воспринимает опровергающую критику. Настоящая работа посвящена моделированию данного процесса.

Наиболее ранние модели слухов (заметим, что, вообще говоря, в данной области термин «слух» не обязательно означает, что информация не подтверждена), не учитывающие влияние средств массовой информации на социум и рассматривающие лишь распространение информации при межличностной коммуникации, были предложены еще в 1964 и 1973 годах [1, 2]. Эти работы породили довольно большое направление в моделировании. Например, модель, в которой численность индивидов изменяется со временем, изучалась в [3], а модель с несколькими группами распространителей информации рассматривалась в [4]. Это лишь некоторые из огромного количества работ, которые разрабатывают подходы, предложенные в [1, 2].

Однако в плане перспективы приложения к реальным социальным процессам эти подходы содержат серьезные недостатки. Как показано в [5], они приводят к гротескному выводу, что все слухи (которые когда-либо были в любую историческую эпоху и в любом обществе) охватывают ту же долю населения к концу своей циркуляции. Этот гротескный вывод возникает из-за того, что указанные модели основаны на предположении, что единственной (и неизбежной) причиной прекращения распространения слуха является то, что когда два его распространителя встречаются, то они приходят к мысли, что этот слух больше не является новостью, и его больше не стоит распространять. Другими словами, в данных подходах свойство конкретного слуха быть более или менее интересным не

имеет никакого отношения к его угасанию. Таким образом, данные модели основаны на неадекватных представлениях о механизме прекращения циркуляции слухов и вряд ли могут претендовать на описание социальной реальности.

Альтернативный подход к моделированию распространения информации в социуме был предложен в [6] и развит в ряде публикаций, включающем [7, 8]; он лежит и в основе настоящей работы.

Среди других направлений данной тематики отметим моделирование информационного влияния и динамики мнений пользователей социальных сетей [9–14], а также анализ лингвистических аспектов онлайновой коммуникации (см., например, [15–18]).

МОДЕЛЬ С НЕПРЕРЫВНЫМ ВРЕМЕНЕМ

Рассмотрим следующую ситуацию. В обществе распространяется ложный слух. Это распространение происходит путем межличностной коммуникации: люди пересказывают его друг другу. В то же время СМИ распространяют информацию, делающую индивидов, незнакомых со слухом, невосприимчивыми к нему. Однако человек, знакомый со слухом, не разубеждается в нем. При этом опровергающая информация, распространяемая СМИ, не рассматривается индивидами как интересная, и они ее не пересказывают.

Таким образом, данная модель предполагает три категории индивидов.

Восприимчивые – те, кто не знаком ни со слухом, ни с его опровержением.

Адепты – знакомые со слухом, верящие в его истинность и распространяющие его.

Инокулированные – знакомые с опровержением слуха и вследствие этого невосприимчивые к нему.

Обозначив численность социальной группы через N, а численности адептов и инокулированных в момент времени t соответственно через $x(t)$ и $y(t)$, получим уравнения вида

$$\frac{dx}{dt} = \beta^* x\big(N - x - y\big),$$
$$\frac{dy}{dt} = \alpha\big(N - x - y\big).$$

Здесь предполагается, что для скорость увеличения численности адептов пропорциональна произведению численности адептов (т. е. тех, кто пересказывает инокулированных) на численность восприимчивых, т. е. $(N-x-y)$. Скорость увеличения численности пропорциональна численности восприимчивых.

Параметры β^*, α описывают интенсивности этих процессов. Из соображений размерности примем $\beta^* = \beta / N$.

Таким образом, уравнения модели принимают вид

$$\frac{dx}{dt} = \frac{\beta}{N} x(N-x-y), \tag{1}$$

$$\frac{dy}{dt} = \alpha(N-x-y). \tag{2}$$

Начальное значение численности адептов положим равным единице. Другими словами, будем считать, что изначально слух распространяется ровно одним индивидом:

$$x(0) = 1. \tag{3}$$

Начальное значение численности инокулированных равно нулю:

$$y(0) = 0. \tag{4}$$

Итак, модель имеет вид (1)–(4).

Очевидно, $dx/dt > 0, dy/dt > 0$ при $0 \le x + y < N$. Таким образом, с течением времени численности адептов и инокулированных возрастают, а численность восприимчивых уменьшается до тех пор, пока эта категория не будет исчерпана. Основной вопрос в данном случае состоит в нахождении соотношения между финальными численностями адептов и инокулированных. Чтобы найти это соотношение, поделим (2) на (1), получим

$$\frac{dy}{dx} = \frac{\alpha N}{\beta x}.$$

Отсюда следует $y = \frac{\alpha N}{\beta} \ln x + C$, где константа интегрирования *C* определяется из условий (3), (4). Подставив $y = 0, x = 1$, получим $C = 0$. Таким образом, связь между численностями адептов и инокулированных в любой момент времени имеет вид

$$y(t) = \frac{\alpha N}{\beta} \ln x(t). \tag{5}$$

Когда численность восприимчивых обращается в нуль, соответствующие численности связаны соотношением

$$x_{fin} + y_{fin} = N . \quad (6)$$

Учитывая (5), получаем

$$x_{fin} + \frac{\alpha N}{\beta} \ln x_{fin} = N . \quad (7)$$

Итак, конечная численность адептов слуха находится из уравнения (7), а конечная численность инокулированных – из уравнения (6).

МОДЕЛЬ С ДИСКРЕТНЫМ ВРЕМЕНЕМ

Чтобы приблизить модель к возможности оценивать реальные процессы, рассмотрим ее дискретный вариант. В качестве единицы времени примем один день. Уравнения модели принимают форму

$$x_{t+1} = x_t + \frac{b}{N} x_t \left(N - x_t - y_t \right), \quad t \geq 1, \quad (8)$$

$$y_{t+1} = y_t + a\left(N - x_t - y_t \right), \quad t \geq 1. \quad (9)$$

Начальные условия имеют вид

$$x_1 = 1, \quad (10)$$

$$y_1 = 0 . \quad (11)$$

В уравнения (8), (9) необходимо внести корректировку для последнего дня распространения слуха. Эта необходимость возникает в связи с тем, что ввиду дискретного характера модели сумма численностей адептов и инокулированных, вычисленных по формулам (8), (9), может превысить численность социума (в случае модели с непрерывным временем этого не происходит, так как сумма $x(t) + y(t)$ приближается к численности социума асимптотически).

Внесем эту корректировку так, чтобы обеспечить пропорциональность приростов $\left(x_{t+1} - x_t \right), \left(y_{t+1} - y_t \right)$. Именно, пусть для вычисленных по формулам (8), (9) значений переменных имеем $x_t + y_t < N$, $x_{t+1} + y_{t+1} > N$. Тогда положим

$$x_{fin} = x_t + \frac{\left(x_{t+1} - x_t \right)}{\left(x_{t+1} - x_t \right) + \left(y_{t+1} - y_t \right)} \left(N - x_t - y_t \right), \quad (12)$$

$$y_{fin} = y_t + \frac{(y_{t+1} - y_t)}{(x_{t+1} - x_t) + (y_{t+1} - y_t)}(N - x_t - y_t). \quad (13)$$

Если же для некоторого значения времени t имеем $x_t + y_t = N$, то положим $x_{fin} = x_t$, $y_{fin} = y_t$, что формально также описывается формулами (12), (13). Таким образом, модель имеет вид (8)–(13).

Значения x_{fin}, y_{fin} представляют основной предмет внимания при моделировании. Вопрос состоит в том, насколько эффективным является опровержение слухов с помощью СМИ.

Параметры b, a в уравнениях (8), (9) имеют смысл, аналогичный параметрам α, β. Для их оценки воспользуемся следующими соображениями. Рассмотрим момент времени t=0, когда слух распространяется ровно одним индивидом (его зачинателем). Имеем

$$x_2 = 1 + b\frac{N-1}{N}.$$

Учитывая, что реальные социумы исчисляются как минимум сотнями тысяч человек, пренебрежем отличием множителя $(N-1)/N$ от единицы. Получим, что параметр b имеет смысл количества индивидов, которым один индивид передает слух за единицу времени при условии, что все контакты являются восприимчивыми к этому слуху (т. е. среди них нет индивидов, уже знакомых с этим слухом, и нет инокулированных). Для вычислительных экспериментов с моделью (8)–(13) будут приняты значения $b = 7$ и $b = 4$.

Параметр a, очевидно, имеет смысл доли восприимчивых, переходящих в категорию инокулированных за единицу времени. Таким образом, он характеризует влияние СМИ. Для телевизионных программ этот параметр можно оценить с помощью рейтинга, под которым понимается «среднее количество человек, смотревших телеканал/телепрограмму, выраженное в процентах от населения (в рамках выбранной ЦА)» (целевой аудитории) [19].

Так, по данным компании Mediascope на неделе 9–15 ноября 2020 года самыми популярными (Россия: города с населением 100 тысяч человек и более) телепрограммами жанра «Информационно-аналитические передачи (комментарии)» были:

- Итоги недели с Ирадой Зейналовой (НТВ; 2020-11-15 рейтинг=4,6).

- Вести недели (Россия 1; 2020-11-15 рейтинг=3,9).

- Постскриптум (ТВЦ; 2020-11-14 рейтинг=2,3).

Укажем также данные нескольких программ жанра «Социально-политические программы». Перечисленные ниже передачи не обязательно являются лидерами рейтинга в данной категории, но приведены нами ввиду их фокусированности на политических вопросах (в указанный жанр входят также такие не-политические программы, как «Человек и закон», «Наш потреб надзор» и т.д.).

- 60 минут (Россия 1; 2020-11-12 рейтинг=3,6).

- Воскресный вечер с Владимиром Соловьевым (Россия 1; 2020-11-15 рейтинг=2,3).

- Вечер с Владимиром Соловьевым (Россия 1; 2020-11-10 рейтинг=2,2).

- Место встречи (НТВ; 2020-11-10 рейтинг=1,8).

С опорой на приведенные значения рейтингов для численных экспериментов с моделью (8)–(13) были приняты значения параметра a порядка нескольких процентов.

ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ: ОПРОВЕРЖЕНИЕ НАЧИНАЕТСЯ В ДЕНЬ ПОЯВЛЕНИЯ ПЕРВОГО АДЕПТА

Эксперимент 1. Положим $N = 10^7$ (10 млн. чел.), $a = 0,02; b = 7$. Результаты расчета представлены на рис. 1. Для данных значений параметров численность адептов ложного слуха оказалась в несколько раз больше численности инокулированных. Очевидно, это происходит потому, что первая из этих величин возрастает почти экспоненциально, а вторая – линейно.

Эксперимент 2. Положим $N = 10^7$ (10 млн. чел.), $a = 0,06; b = 7$. Результаты расчета представлены на рис. 2. Значение параметра $a = 0,06$ соответствует ситуации, когда опровержение слуха проводится различными СМИ. Арифметическая сумма рейтингов, например, программ «Итоги недели с Ирадой Зейналовой» и «60 минут» превосходит это значение (4,6+3,6=8,2>6). Однако с учетом того, что аудитории этих

программ частично пересекаются, а также того, что рейтинги ежедневных программ ниже, чем у еженедельных, представляется, что оно может быть принято для расчета оценки численностей адептов и инокулированных. Результаты показывают, что даже совместное действие различных СМИ не позволяет инокулировать даже половину социума.

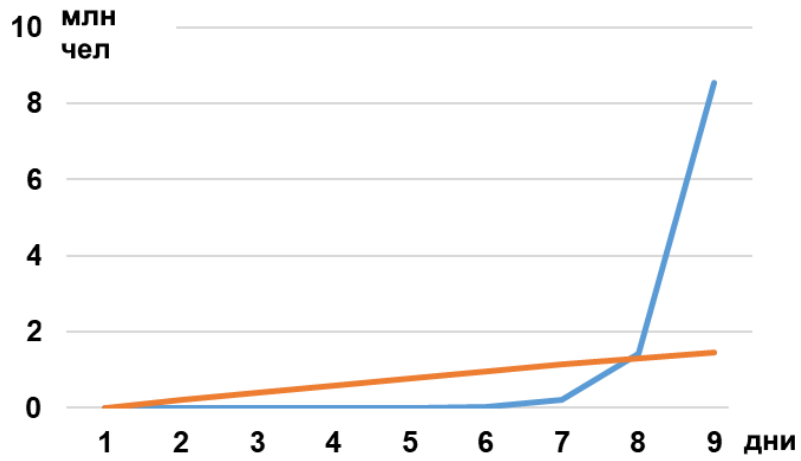


Рис. 1. Эксперимент 1: численности адептов $x(t)$ (синяя линия) и инокулированных $y(t)$ (оранжевая)

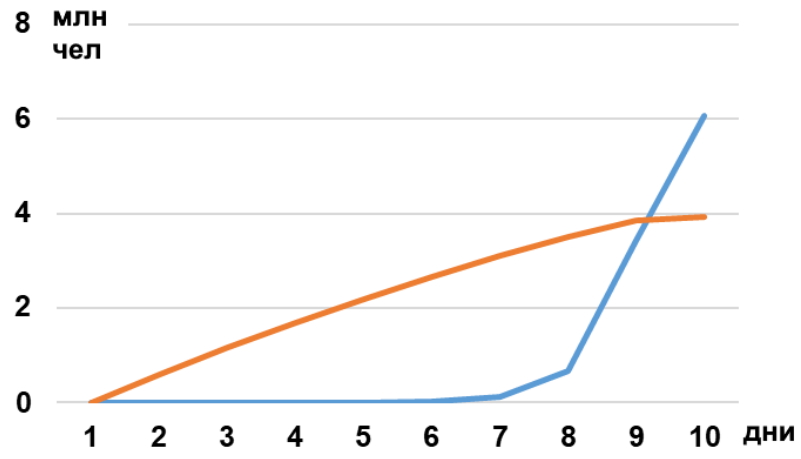


Рис. 2. Эксперимент 2: численности адептов $x(t)$ (синяя линия) и инокулированных $y(t)$ (оранжевая)

Эксперимент 3. Положим $N = 10^7$ (10 млн. чел.), $a = 0{,}06; b = 4$. Результаты расчета представлены на рис. 3. В данном случае слух является менее вирусным, и опровергающим его СМИ удается добиться примерного паритета.

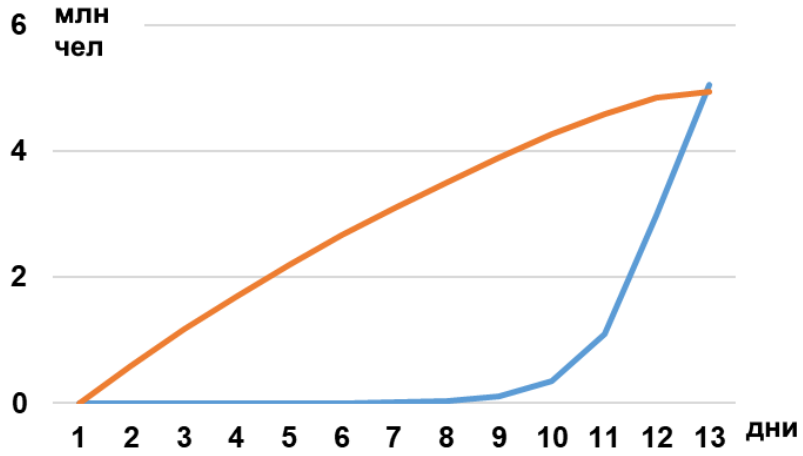


Рис. 3. Эксперимент 3: численности адептов $x(t)$ (синяя линия) и инокулированных $y(t)$ (оранжевая)

СЛУЧАЙ ЗАПАЗДЫВАЮЩЕГО ОПРОВЕРЖЕНИЯ

В приведенных выше экспериментах опровергающее вещание средств массовой информации начиналось в тот же день, когда начиналось распространение слуха. Рассмотрим теперь случай, когда вещание начинается с запаздыванием. Именно, положим, что

$$y_1 = y_2 = 0, \tag{14}$$

а уравнение (9) действует при $t \geq 2$. Для данной модели проведем вычислительный эксперимент с теми же значениями параметров, что в Эксперименте 3.

Эксперимент 4. Положим $N = 10^7$ (10 млн. чел.), $a = 0{,}06; b = 4$, модель имеет вид (8)–(14). Результаты расчета представлены на рис. 4. Они показывают, что последствия небольшой задержки в реакции не являются драматическими, хотя и приводят к некоторому увеличению численности адептов (по сравнению с Экспериментом 3).

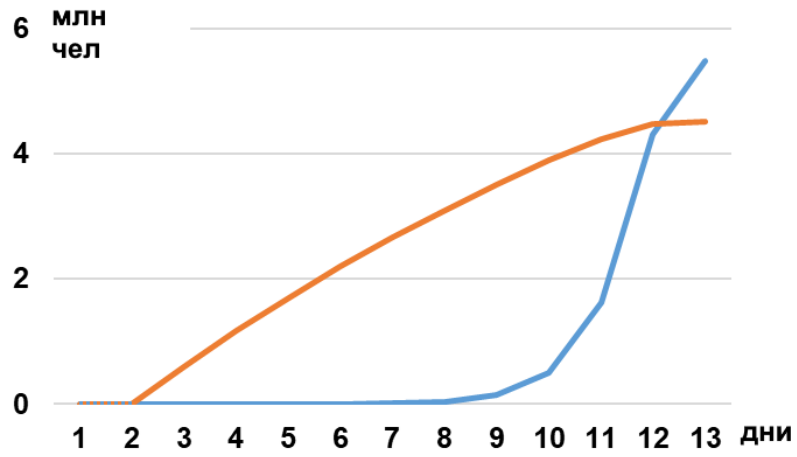


Рис. 4. Эксперимент 4: численности адептов $x(t)$ (синяя линия) и инокулированных $y(t)$ (оранжевая)

ЗАКЛЮЧЕНИЕ

Основной вывод проведенного моделирования состоит в том, что средствам массовой информации трудно противостоять слухам. Причина состоит в том, что при появлении нового вопроса в информационной повестке дня слух на эту тему разрастается на первых порах в режиме, близком к экспоненциальному, а численность людей, получивших информацию от СМИ, – медленнее, чем линейно. Однако эти закономерности были получены в предположении, что информация СМИ не пересказывается индивидами (телезрителями) своим соседям по социуму. Соответственно возможность противостоять слуху состоит в том, чтобы СМИ создавали вирусный контент, способный передаваться при межличностной коммуникации аналогично слуху.

СПИСОК ЛИТЕРАТУРЫ

1. *Daley D.J., Kendall D.G.* Stochastic rumors // Journal of the Institute of Mathematics and its Applications. 1964. V. 1. P. 42–55.

2. *Maki D.P., Thompson M.* Mathematical Models and Applications. Prentice-Hall, Englewood Cliffs, NJ, USA, 1973.

3. *Chen Guanghua, Shen H., Ye T., Chen G., Kerr N.* A kinetic model for the spread of rumor in emergencies // Discrete dynamics in nature and society. 2013. V. 2013. Article ID 605854. 8 p.

4. *Isea R., Mayo-García R.* Mathematical analysis of the spreading of a rumor among different subgroups of spreaders // Pure and Applied Mathematics Letters. 2015. V. 2015. P. 50–54.

5. *Mikhailov A.P., Pronchev G.B., Proncheva O.G.* Mathematical Modeling of Information Warfare in Techno-Social Environments // Techno-Social Systems for Modern Economical and Governmental Infrastructures. IGI Global. 2019. P. 174–210.

6. *Самарский А.А., Михайлов А.П.* Математическое моделирование (Идеи, Методы, Примеры), 1997.

7. *Михайлов А.П., Петров А.П., Прончева О.Г.* Модель информационного противоборства в социуме с кусочно-постоянной функцией дестабилизирующего воздействия // Математическое моделирование. 2018. Т. 30, № 7. С. 47–60.

8. *Petrov A.P., Lebedev S.A.* Online Political Flashmob: The Case of 632305222316434 // Computational mathematics and information technologies. 2019. No. 1. P. 17–28. https://doi.org/10.23947/2587-8999-2019-1-1-17-28

9. *Chartishvili A.G., Kozitsin I.V., Goiko V. L., Saifulin E.R.* On an Approach to Measure the Level of Polarization of Individuals' Opinions // 2019 Twelfth International Conference "Management of large-scale system development" (MLSD), Moscow, Russia, 2019. P. 1–5. https://doi.org/10.1109/MLSD.2019.8911015.

10. *Kozitsin I.V., Marchenko A.M., Goiko V.L., Palkin R.V.* Symmetric Convex Mechanism of Opinion Formation Predicts Directions of Users' Opinions Trajectories // 2019 Twelfth International Conference "Management of large-scale system development" (MLSD), Moscow, Russia, 2019. P. 1–5. https://doi.org/10.1109/MLSD.2019.8911064.

11. *Kozitsin I.V., Chkhartishvili A.G., Marchenko A.M., Norkin D.O., Osipov S.D., Uteshev I.A., Goiko V.L., Palkin R.V., Myagkov M.G.* Modeling Political Preferences of Russian Users Exemplified by the Social Network Vkontakte // Mathematical Models and Computer Simulations. 2020. V. 12. P. 185–194.
https://doi.org/10.1134/S2070048220020088.

12. *Chkhartishvili A.G., Gubanov D.A., Novikov D.A.* Social Networks: Models of information influence, control and confrontation. Cham, Switzerland: Springer International Publishing, 2019. 158 p. https://doi.org/10.1007/978-3-030-05429-8.

13. *Губанов Д.А., Чхартишвили А.Г.* Влиятельность пользователей и метапользователей социальной сети // Проблемы управления. 2016. № 6. С. 12–17.

14. *Chkhartishvili A.G, Gubanov D.A.* On Approaches to Identifying Information Spread Channels in Online Social Networks // 2019 Twelfth International Conference "Management of large-scale system development" (MLSD), Moscow, Russia, 2019. P. 1–5. https://doi.org/10.1109/MLSD.2019.8911065

15. *Akhtyamova L., Alexandrov M., Cardiff J., Koshulko O.* Opinion Mining on Small and Noisy Samples of Health-related Texts // Advances in Intelligent Systems and Computing III (Proc. of CSIT-2018), Springer, AISC. 2019. V. 871. P. 1–12.

16. *Akhtyamova L., Cardiff J.* LM-Based Word Embeddings Improve Biomedical Named Entity Recognition: A Detailed Analysis // Bioinformatics and Biomedical Engineering. IWBBIO 2020. Lecture Notes in Computer Science. Springer, Cham, 2020. V. 12108. https://doi.org/10.1007/978-3-030-45385-5_56

17. *Boldyreva A., Sobolevskiy O., Alexandrov M., Danilova V.* Creating collections of descriptors of events and processes based on Internet queries // Proc. of 14-th Mexican Intern. Conf. on Artif. Intell. (MICAI-2016), Springer Cham, LNAI, 2016. Vol. 10061 (chapter 26). P. 303–314.

18. *Boldyreva A., Alexandrov M., Koshulko O., Sobolevskiy O.* Queries to Internet as a tool for analysis of the regional police work and forecast of the crimes in regions // Proc. of 14-th Mexican Intern. Conf. on Artif. Intell. (MICAI-2016), Springer Cham, LNAI, 2016. V. 10061 (chapter 25). P. 290–302.

19. Mediascope. URL: https://mediascope.net/data/#popup_definition_tv

REFUTATION OF A RUMOR BY THE MASS MEDIA: MATHEMATICAL MODEL AND NUMERICAL EXPERIMENTS

A. P. Mikhailov[1, [0000-0002-2730-1538]], A. P. Petrov[2, [0000-0001-5244-8286]]

Keldysh Institute of Applied Mathematics, Moscow

apmikhailov@yandex.ru, petrov.alexander.p@yandex.ru

Abstract

The process is considered, in which an unreliable rumor spreads in society, which is opposed by the broadcasting of the mass media. In this case, the unreliability of hearing is understood so that the information of the media contains a refutation and thereby inoculates individuals, that is, makes them immune to hearing. At the same time, individuals who have managed to accept the rumor cease to trust the media and thereby become unavailable for persuasion. For this process, a mathematical model is proposed in two versions. The variant with continuous time reveals some of the mathematical properties of the model. The discrete time option is more convenient for analyzing real processes since it allows one to estimate the parameters of the model. To assess these parameters, data on the ratings of the main socio-political programs of Russian TV channels were used. Several scenario calculations of the model with these parameters are presented. The main conclusion is that if the information disseminated by the media is not viral, that is, it is not retold by viewers to their neighbors in society, then the media are unable to resist rumors.

***Keywords**: mathematical modeling, information warfare, numerical experiment, rumors.*

REFERENCES

1. *Daley D.J., Kendall D.G.* Stochastic rumors // Journal of the Institute of Mathematics and its Applications. 1964. V. 1. P. 42–55.

2. *Maki D.P., Thompson M.* Mathematical Models and Applications. Prentice-Hall, Englewood Cliffs, NJ, USA, 1973.

3. *Chen Guanghua, Shen H., Ye T., Chen G., Kerr N.* A kinetic model for the spread of rumor in emergencies // Discrete dynamics in nature and society. 2013. V. 2013. Article ID 605854. 8 p.

4. *Isea R., Mayo-García R.* Mathematical analysis of the spreading of a rumor among different subgroups of spreaders // Pure and Applied Mathematics Letters. 2015. V. 2015. P. 50–54.

5. *Mikhailov A.P., Pronchev G.B., Proncheva O.G.* Mathematical Modeling of Information Warfare in Techno-Social Environments // Techno-Social Systems for Modern Economical and Governmental Infrastructures. IGI Global. 2019. P. 174–210.

6. *Samarskii A.A., Mikhailov A.P.* Principles of Mathematical Modelling: Ideas, Methods, Examples. Taylor and Francis Group, 2001.

7. *Mikhailov A.P., Petrov A.P., Proncheva O.G.* A Model of Information Warfare in a Society with a Piecewise Constant Function of the Destabilizing Impact // Mathematical Models and Computer Simulations. 2019. V. 11, No. 2. P. 190–197.

8. *Petrov A.P., Lebedev S.A.* Online Political Flashmob: The Case of 632305222316434 // Computational mathematics and information technologies. 2019. No. 1. P. 17–28. https://doi.org/10.23947/2587-8999-2019-1-1-17-28

9. *Chartishvili A.G., Kozitsin I.V., Goiko V.L., Saifulin E.R.* On an Approach to Measure the Level of Polarization of Individuals' Opinions // 2019 Twelfth International Conference "Management of large-scale system development" (MLSD), Moscow, Russia, 2019. P. 1–5. https://doi.org/10.1109/MLSD.2019.8911015.

10. *Kozitsin I.V., Marchenko A.M., Goiko V.L., Palkin R.V.* Symmetric Convex Mechanism of Opinion Formation Predicts Directions of Users' Opinions Trajectories // 2019 Twelfth International Conference "Management of large-scale system development" (MLSD), Moscow, Russia, 2019. P. 1-5. https://doi.org/10.1109/MLSD.2019.8911064.

11. *Kozitsin I.V., Chkhartishvili A.G., Marchenko A.M., Norkin D.O., Osipov S.D., Uteshev I.A., Goiko V.L., Palkin R.V, Myagkov M. G.* Modeling Political Preferences of Russian Users Exemplified by the Social Network Vkontakte // Mathematical Models and Computer Simulations. 2020. V. 12. P. 185–194.
https://doi.org/10.1134/S2070048220020088.

12. *Chkhartishvili A.G., Gubanov D.A., Novikov D.A.* Social Networks: Models of information influence, control and confrontation. Cham, Switzerland: Springer International Publishing, 2019. 158 p. https://doi.org/10.1007/978-3-030-05429-8.

13. *Chkhartishvili A.G, Gubanov D.A.* Influence Levels of Users and Meta-Users of a Social Network // Automation and Remote Control. 2018. V. 79, Issue 3. P. 545–553, https://doi.org/10.1134/S0005117918030128

14. *Chkhartishvili A.G, Gubanov D.A.* On Approaches to Identifying Information Spread Channels in Online Social Networks // 2019 Twelfth International Conference "Management of large-scale system development" (MLSD), Moscow, Russia, 2019. P. 1–5. https://doi.org/10.1109/MLSD.2019.8911065

15. *Akhtyamova L., Alexandrov M., Cardiff J., Koshulko O.* Opinion Mining on Small and Noisy Samples of Health-related Texts // Advances in Intelligent Systems and Computing III (Proc. of CSIT-2018), Springer, AISC. 2019. V. 871. P. 1–12.

16. *Akhtyamova L., Cardiff J.* LM-Based Word Embeddings Improve Biomedical Named Entity Recognition: A Detailed Analysis // Bioinformatics and Biomedical Engineering. IWBBIO 2020. Lecture Notes in Computer Science. Springer, Cham, 2020. V. 12108. https://doi.org/10.1007/978-3-030-45385-5_56

17. *Boldyreva A., Sobolevskiy O., Alexandrov M., Danilova V.* Creating collections of descriptors of events and processes based on Internet queries // Proc. of 14-th Mexican Intern. Conf. on Artif. Intell. (MICAI-2016), Springer Cham, LNAI, 2016. V. 10061 (chapter 26). P. 303–314.

18. *Boldyreva A., Alexandrov M., Koshulko O., Sobolevskiy O.* Queries to Internet as a tool for analysis of the regional police work and forecast of the crimes in regions // Proc. of 14-th Mexican Intern. Conf. on Artif. Intell. (MICAI-2016), Springer Cham, LNAI, 2016. V. 10061 (chapter 25). P. 290–302.

19. Mediascope. URL: https://mediascope.net/data/#popup_definition_tv

СВЕДЕНИЯ ОБ АВТОРАХ

МИХАЙЛОВ Александр Петрович – д. ф.-м. н., главный научный сотрудник Института прикладной математики им. М.В. Келдыша РАН;

Alexander Petrovich MIKHAILOV – D. Sci in Applied Mathematics, Chief Researcher at KIAM.

apmikhailov@yandex.ru

ПЕТРОВ Александр Пхоун Чжо – д. ф.-м. н., ведущий научный сотрудник Института прикладной математики им. М.В. Келдыша РАН.

Alexander PETROV – D. Sci in Applied Mathematics, Leading Researcher at KIAM. Research area: mathematical modeling in social science.

email: petrov.alexander.p@yandex.ru

Материал поступил в редакцию 25 ноября 2020 года

УДК 002.01

ПРЕПРИНТ КАК МАТЕРИАЛ ДЛЯ ОВЕРЛЕЙНОГО ЖУРНАЛА

Т. А. Полилова[0000-0003-4628-3205]

Федеральное государственное учреждение «Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук»

polilova@keldysh.ru

Аннотация

Движение Открытого доступа имеет давнюю историю. В 2002 г. впервые была озвучена Будапештская инициатива Открытого доступа. Однако до сих пор проблема Открытого доступа к научным публикациям не получила своего полного и окончательного решения. В 2018 г. в Европейском союзе был принят План S, который предписывает к 2020 г. сделать открытый доступ реальностью. План S подчеркивает важность самоархивирования статей и роль архивов (серверов) препринтов для размещения научных результатов. Отмечается, что архивы препринтов обладают большим потенциалом для редакционно-издательских инноваций. Научные журналы ограниченного для читателя доступа, функционирующие на коммерческой основе, не сдают своих позиций. Но и здесь мы видим определенные подвижки. Журналы стали менее жестко формулировать свою политику по отношению к препринтам и постпринтам статей.

Все больше зарубежных ученых становятся приверженцами движения «Справедливый открытый доступ», которое предлагает новое организационное решение. Журнал должен иметь учредителя в лице научной организации или некоммерческого фонда, которые нанимают группу исполнителей на оказание редакционно-издательских услуг. Редакторы и издатели не должны иметь своих коммерческих интересов. Финансирование научного журнала должно осуществляться за счет общего вклада организаций.

В статье рассматривается современный тип онлайнового научного журнала – оверлейный журнал. Себестоимость выпуска оверлейного журнала настолько низкая, что журнал легко может реализовать схему «бесплатно для автора, бесплатно для читателя». Оверлейный журнал опирается на общедоступные архивы

(серверы) препринтов. Оверлейный онлайновый журнал проводит рецензирование статьи, поступившей из архива, в случае принятия статьи к публикации размещает на своем сайте ее метаданные, а сама скорректированная статья (ее полный текст) вновь размещается в архиве. Такая схема работы не перегружает функциональность архива, но при этом позволяет снизить финансовую нагрузку на оверлейный журнал.

***Ключевые слова**: научный журнал, Справедливый Открытый доступ, Открытый архив, сервер препринтов, оверлейный журнал.*

ВВЕДЕНИЕ

Представление о том, что результаты научных исследований должны находиться в свободном доступе, близко и понятно ученым [1]. Речь в первую очередь идет об опубликовании результатов исследований, выполняемых за счет бюджетных средств и средств публичных научных фондов. В обществе сложилось понимание, что Открытый доступ к научным достижениям помогает развитию науки и технологий, способствует техническому и гуманитарному прогрессу.

Движение Открытого доступа имеет давнюю историю. В 2002 г. впервые была озвучена Будапештская инициатива Открытого доступа. Открытый доступ определяется как свободный доступ к научной информации через общедоступный интернет, предполагающий право каждого пользователя читать, загружать, копировать, распространять или использовать для других законных целей научные публикации при отсутствии финансовых, правовых и технических преград, за исключением тех, которые регулируют доступ собственно к интернету. В 2003 г. принята Берлинская декларация по Открытому доступу. Усилиями сторонников Открытого доступа за прошедшие годы заметно изменился ландшафт научных публикаций. Стали массово появляться журналы Открытого доступа, которые составили серьезную конкуренцию традиционным коммерческим научным журналам. Однако до сих пор проблема Открытого доступа к научным публикациям не получила своего полного и окончательного решения. В 2018 г. в Европейском союзе был принят План S, который предписывает к 2020 г. сделать открытый доступ реальностью. План S предлагает конкретные шаги для внедрения Открытого доступа в широкую практику научных журналов.

Отметим, что План S подчеркивает важность самоархивирования статей и роль общедоступных серверов препринтов для размещения научных результатов. Отмечается, что архивы препринтов обладают большим потенциалом для редакционно-издательских инноваций.

Научные журналы ограниченного для читателя доступа, функционирующие на коммерческой основе, не сдают своих позиций. Но и здесь мы видим определенные подвижки. Журналы стали менее жестко формулировать свою политику по отношению к препринтам и постпринтам статей. Так, например, журналы издательства Springer допускают предварительное размещение препринтов в открытых институциональных архивах.

МЕСТО ПРЕПРИНТОВ В ИНФРАСТРУКТУРЕ НАУЧНЫХ ПУБЛИКАЦИЙ

Что представляет собой сегодняшний препринт? Восприятие препринта как издания, не прошедшего рецензирования, уходит в прошлое. Более точно – препринту возвращается то доверие, которое существовало в 1950–1960-е годы. В те годы многие институты Академии наук выпускали печатные препринты, подготовленные сотрудниками институтов. Препринты были востребованы среди ученых. В ИПМ им. М.В. Келдыша РАН, например, выпускались небольшие, в несколько десятков экземпляров, тиражи препринтов, которые передавались в академические институты и Книжную палату для формирования фондов научных библиотек. Препринты нашего Института перед публикацией обсуждались на семинарах, проходили строгую оценку руководителей и ведущих научных сотрудников. В виде препринтов были опубликованы многие фундаментальные результаты работ ученых.

Некоторые издатели не считают препринт серьезным изданием, поскольку препринт не проходит процедуру рецензирования, принятую в журналах. Однако препринту не противопоказаны рецензирование или модерация. Возможно, в архиве препринтов можно встретить недостаточно обоснованную или не слишком ответственную статью. Однако следует признать, что то же самое можно сказать и о некоторых статьях, прошедших рецензирование и опубликованных в престижных журналах. Ни одна система, управляемая людьми, никогда не будет безупречной, и экспертная оценка не исключение [2].

Препринты чаще всего публикуются на сайтах научных организаций и вузов, в тематических архивах препринтов. Авторитет препринтов вырос благодаря успешной деятельности таких архивов препринтов, как arXiv, BiorXiv, MedRxiv. В этих архивах в обязательном порядке проводится модерация поступающих статей: эксперты проверяют научный уровень статьи, актуальность статьи для предметной области. Перечисленные архивы препринтов обладают развитой функциональностью. Последнее время к архиву препринтов такого уровня стали применять термин «сервер препринтов».

В ИПМ им. М.В. Келдыша РАН препринты превратились в полноценное рецензируемое научное издание [3], где каждый препринт проходит внутреннее рецензирование в подразделении Института, в котором авторы препринтов проводили свои исследования. Ответственность за качество научного содержания препринта берет на себя руководитель подразделения, организующий рецензирование и подписывающий рецензию. В случае необходимости к рецензированию привлекаются внешние рецензенты. Научное издание «Препринты ИПМ» имеет ISSN для печатной и онлайновой версий, индексируется в Российском индексе научного цитирования (РИНЦ), входит в Перечень журналов ВАК.

Главное назначение препринта – оперативная публикация результатов научного исследования. Время публикации составляет обычно всего 3–4 дня после получения редакцией материалов. Традиционные журналы в такие сроки не укладываются. Неудивительно, что роль препринтов возросла во время пандемии коронавируса COVID-19. В России в период пандемии создан открытый архив препринтов COVID-19 PREPRINTS (https://covid19-preprints.microbe.ru/) (рис. 1). В проекте участвует Российский научно-исследовательский противочумный институт «Микроб» Роспотребнадзора, техническую поддержку архива препринтов осуществляет НЭИКОН.

Цель разработки COVID-19 PREPRINTS – повысить доступность научных результатов исследований по COVID-19, расширить сотрудничество между исследователями, информировать о текущих исследованиях посредством своевременной онлайновой отчетности. Каждый препринт, размещаемый в архиве, получает DOI.

Читатели могут оставлять публичные комментарии к статьям на COVID-19 PREPRINTS. Комментарии модерируются. Читатели также имеют возможность связаться с авторами по адресам электронной почты, указанным в препринте.

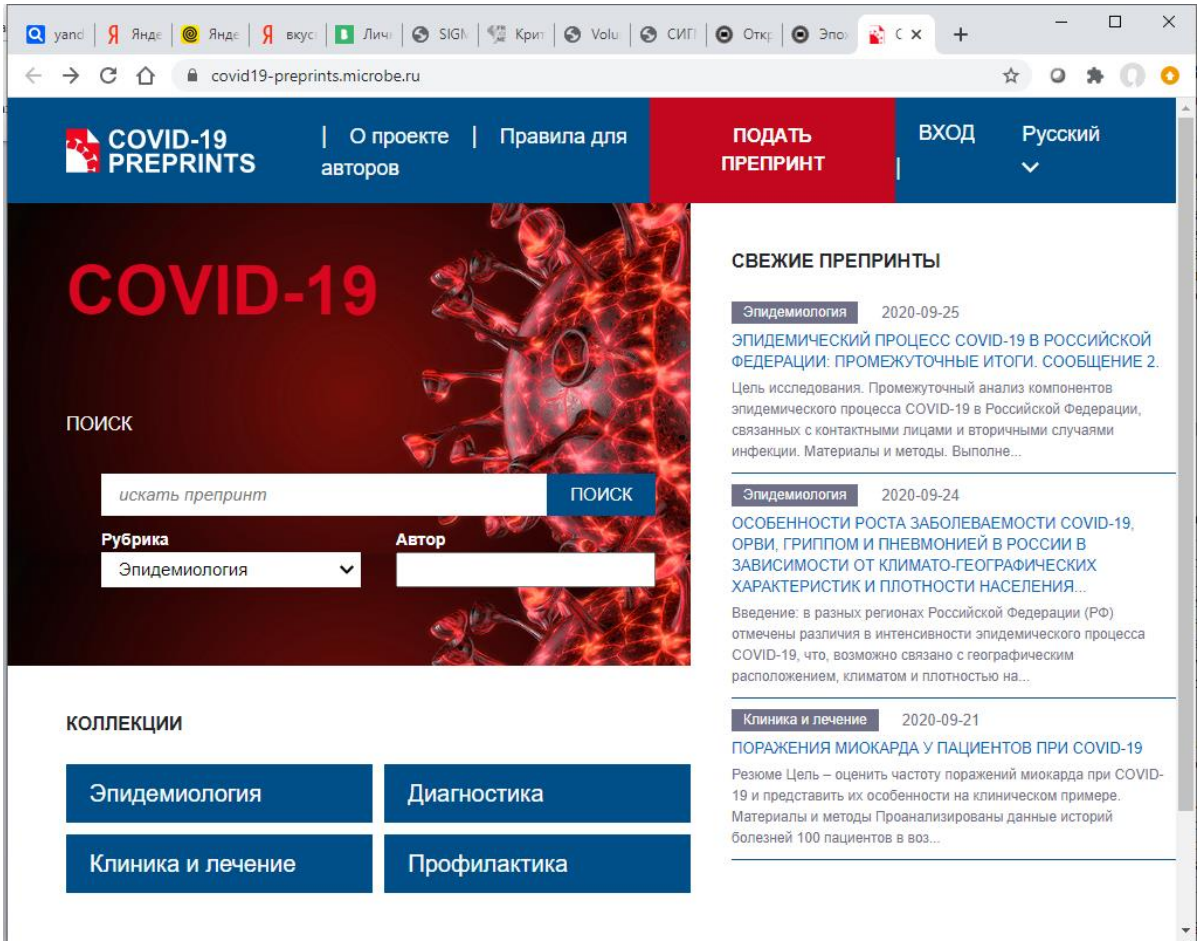


Рис. 1. Сайт проекта COVID-19 PREPRINTS

Материалы препринтов в дальнейшем могут быть опубликованы в журналах. В России Гражданский кодекс РФ защищает права авторов и правообладателей, в качестве которых чаще всего выступают научные организации. Так, например, заключение авторского лицензионного договора с издателем не влечет за собой переход исключительного права на произведение к издателю. Даже если работодатель поручил автору заключить авторский договор с издателем на условиях исключительной лицензии, работодатель всегда сохраняет за собой право опубликовать статью, отосланную в журнал, на своем сайте на условиях простой (неисключительной) лицензии [4–6].

Если журнал, опубликовавший статью по материалам препринта, является изданием ограниченного для читателя доступа, препринт остается открытой версией статьи, доступной широкому читателю.

О ФИНАНСОВОЙ СТОРОНЕ ОТКРЫТОГО ДОСТУПА

Сейчас сосуществуют две альтернативные модели доступа к материалам научных журналов: Открытый доступ и ограниченный доступ (платный доступ к статьям, платная подписка на журнал). Журналы ограниченного доступа и утвердившиеся на рынке издательские дома на Западе не сдают своих позиций. На стороне Открытого доступа – многочисленные энтузиасты, поддержка со стороны научных фондов, некоммерческих организаций и правительственных инициатив [7]. Однако отлаженные и устойчивые финансовые механизмы, позволяющие поступательно развиваться архивам Открытого доступа, во многих странах пока не созданы.

Журналы Открытого доступа предоставляют своим читателям бесплатный доступ к материалам журнала. Одна из бизнес-моделей журналов Открытого доступа предполагает, что расходы на редакционно-издательскую подготовку журнала возмещаются за счет взносов авторов статей. За публикацию статьи автору будет предложено заплатить несколько тысяч долларов США. В то же время есть и другие примеры: за размещение статьи в Открытом архиве препринтов arXiv автор препринта ничего не платит. Себестоимость препринта arXiv оценивается всего в 10 долларов, расходы несет Корнеллский университет (США) [8].

Модель ограниченного доступа к научным журналам реализуют известные западные издательства Springer, Elsevier, Wiley, Informa. Многие ученые справедливо считают, что рецензируемые журналы этих крупных издательств обеспечивают высокие академические стандарты и заслуженно имеют высокие рейтинги. Многие также могут согласиться с тем, что высокое качество журналов неизбежно требует серьезных расходов. По некоторым сведениям [8] публикация одной рецензируемой статьи в журнале Nature обходится в 40 тысяч долларов. Если сравнить эту себестоимость с себестоимостью публикации одного модерируемого препринта в arXiv (10 долларов), то возникает вопрос: какие именно этапы редакционно-издательской подготовки потребовали столь высоких затрат? Возможно,

все намного проще: издатель журнала Nature старается обеспечить себе безбедное существование за счет получения сверхприбыли. Эту сверхприбыль журнал получает за счет сбора денег с читателя через платную подписку и торговлю отдельными статьями.

Если автор решит опубликовать свою статью в западном журнале, то для него финансовый вопрос может встать весьма остро. Автор заинтересован в том, чтобы с его статьей познакомилось как можно большее число читателей. В этом случае он должен был бы выбрать журнал Открытого доступа. Однако взнос порядка 2–5 тысяч долларов США для многих может оказаться непосильным. Некоторые западные научные фонды закладывают в гранты оплату взноса за публикацию статей, имея информацию о существующих тарифах. Но для российского ученого взнос размером в несколько тысяч долларов США оказывается неподъемным: такой взнос трудно покрыть, например, из гранта РФФИ. Поэтому, думая о выборе журнала для публикации статьи, автор может принять решение в пользу журнала ограниченного для читателя доступа без авторского взноса.

СПРАВЕДЛИВЫЙ ОТКРЫТЫЙ ДОСТУП

На Западе все большую популярность приобретает концепция Справедливого Открытого доступа (Fair Open Access). Группа ученых и сотрудников библиотек, объединившихся в альянс Fair Open Access Alliance (FOAA), выступила с резкой критикой коммерциализации издания научных журналов и предлагает новое организационное решение [9].

Идеологи Справедливого Открытого доступа считают, что учредителями журнала должны быть научные учреждения или независимые некоммерческие фонды, которые нанимают группу исполнителей на оказание редакционно-издательских услуг. Редакторы и издатели не должны иметь своих коммерческих интересов. Финансирование научного журнала осуществляется за счет общего вклада университетов, исследовательских организаций, других спонсоров, причем эти вклады не должны быть привязаны к отдельным статьям или группам авторов. Взносы от авторов или их спонсоров не исключаются, но они должны быть добровольными и ненавязчивыми. Отсутствие взноса не может стать причиной

отказа автору в публикации. Также не следует отклонять статью автора, если организация, где работает автор, не является спонсором журнала.

Взнос за публикацию должен быть небольшим, не выше одной тысячи долларов США, лучше – еще более низким. Весь процесс издания и распространения журнала должен быть прозрачным и на любом этапе исключать какие-либо коммерческие интересы. Все затраты должны быть понятными учредителям и спонсорам. Тем самым концепция Справедливого Открытого доступа полностью исключает коммерциализацию научного журнала.

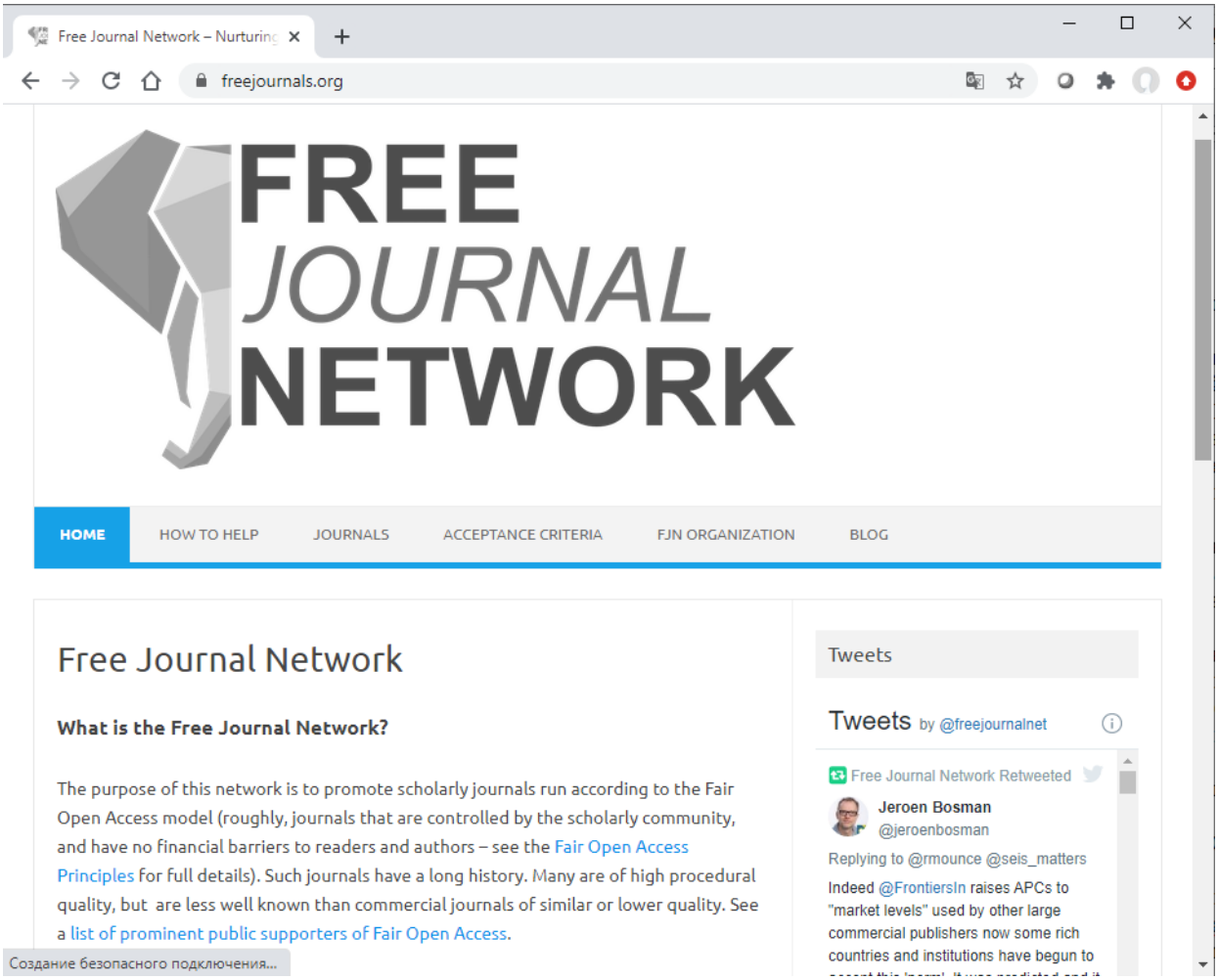


Рис. 2. Сеть бесплатных журналов Free Journal Network

Журналы, поддерживающие принципы Справедливого Открытого доступа, объединяются в ассоциации и сети. Пример такой сети – Сеть бесплатных журналов (Free Journal Network) [10] (рис. 2).

Организаторы сети ставят своей основной задачей помощь журналам в координации усилий по их продвижению, в переходе журналов с коммерческой подпиской на схему Справедливого Открытого доступа. Деятельность сети помогает укрепить экосистему независимых журналов и поставщиков услуг. Сеть координирует свою работу с такими подразделениями FOAA, как LingOA, MathOA, PsyOA, которые сосредоточены на работе с журналами по своим тематическим направлениям. Все эти организации стремятся продемонстрировать, что модель Справедливого Открытого доступа превосходит модель издания коммерческих журналов.

Какова ситуация со Справедливым Открытым доступом в России? Рассмотрим данные сайта крупнейшего российского агрегатора научной периодики eLibrary.ru. На текущий момент в базе агрегатора размещено 14739 российских журналов, выпускаемых на сегодняшний день, из них 6378 – с полными текстами в открытом доступе. На основании этих данных можно заключить, что Открытым доступом в России охвачено около 43% журналов. К сожалению, сведения, опубликованные на сайте eLibrary.ru, не дают возможности оценить, какая доля журналов Открытого доступа взимает с авторов плату за публикацию. Известно, что многие журналы финансируются научными организациями и вузами, и авторы публикуются в этих журналах бесплатно. Но есть журналы, которые предлагают авторам опубликовать статью за скромные деньги. Так, например, размещенный в eLibrary.ru журнал «Столица науки» (рис. 3) готов оперативно опубликовать статью всего за 600 рублей (около 10 долларов США). Это вполне приемлемая цена, укладывающаяся в рекомендации Справедливого Открытого доступа.

Можно было бы сделать оптимистичный вывод о широком внедрении Справедливого Открытого доступа в России, однако на самом деле здесь мы сталкиваемся с другой ситуацией. В то время как на Западе в 1960–1980-х годах набирали силу такие крупные коммерческие издательства, как Springer, Elsevier, Wiley, Informa, в России изданием научных журналов занимались научные институты, вузы, организации с бюджетным финансированием.

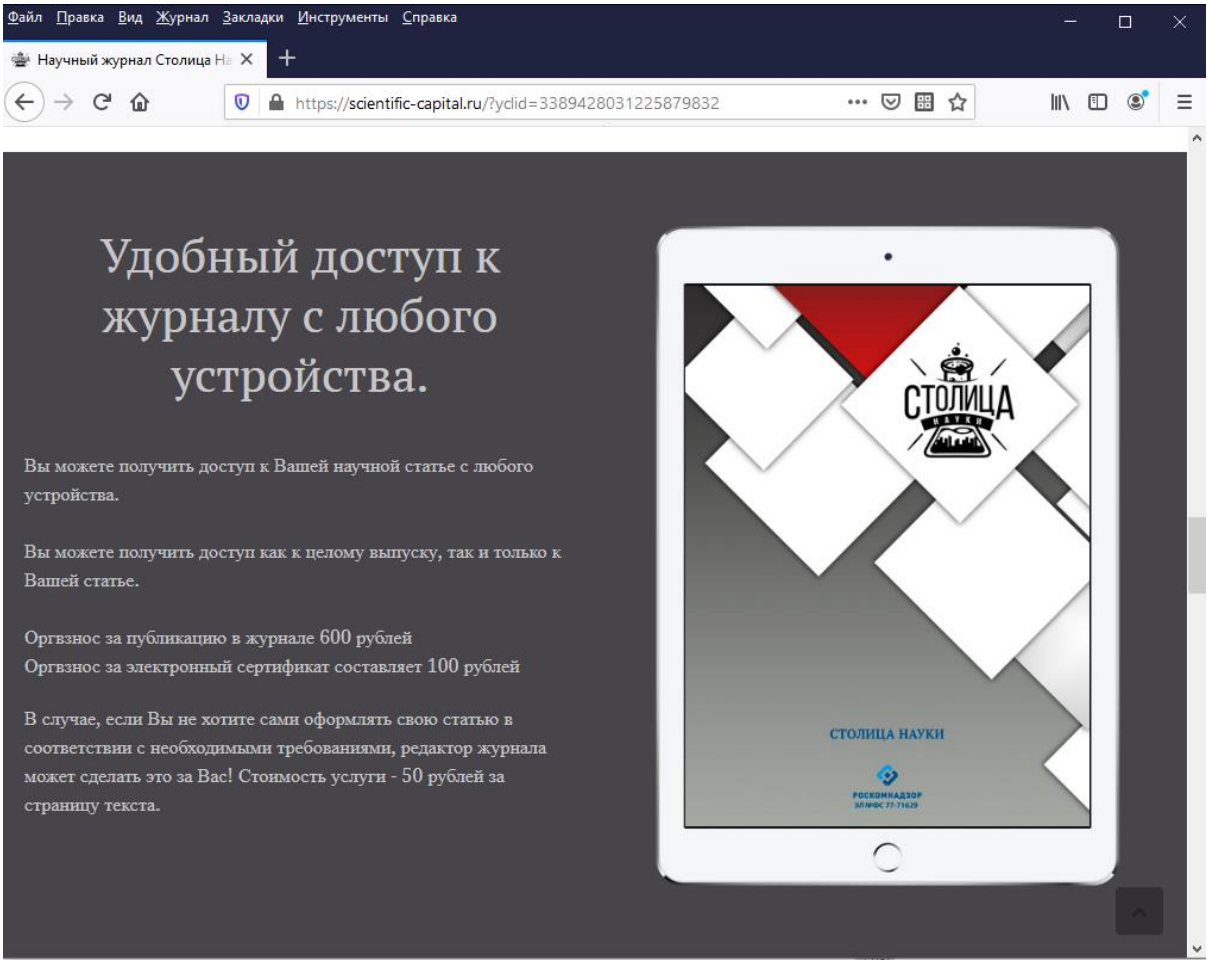


Рис. 3. Предложение о публикации статьи в журнале «Столица науки»

Издание российских научных журналов в то время не рассматривалось как коммерческое предприятие, направленное на получение прибыли. И сейчас в России нет издательств, которые рискнули бы наладить издание научных журналов на коммерческой основе: серьезный бизнес на издании научных журналов в России вряд ли возможен. Таким образом, вполне можно было бы считать, что Справедливый Открытый доступ в нашей стране сформировался еще в 1960–1980-е годы в силу существовавших тогда издательских традиций при государственной финансовой поддержке научной и издательской деятельности. Стоимость подписки на печатные научные журналы была весьма скромной, и подписка не воспринималась научными библиотеками организаций как серьезное бремя. В те годы любой ученый мог позволить себе выписать научный журнал или купить его в свободной продаже.

С появлением интернета российские ученые ожидали, что им будут доступны онлайновые версии российских научных журналов. Действительно, многие научные журналы, созданные на базе научных организаций, стали размещать журнальные выпуски в свободном доступе. Но далеко не все журналы попали в открытый доступ. Например, ведущие академические журналы установили для онлайновых версий журналов платную подписку. Одновременно для академических журналов появилось эмбарго на свободный (бесплатный) доступ длительностью от полугода до трех лет. Ограничение на свободный доступ действует и до сих пор. Стоит отметить, что в 2018 г. такой серьезный контролирующий и надзорный орган, как Счетная Палата РФ, указала на недопустимость ограничений на свободный доступ к академическим журналам, созданным при бюджетном финансировании. После такого замечания академические журналы появились в открытом доступе, но только на несколько месяцев. Сейчас многие академические журналы вновь попадают под эмбарго. Информация о доступе к российским научным математическим (некоторым физическим и естественно-научным) журналам размещена на сайте MathNet (http://www.mathnet.ru/ ej.phtml?option_lang=rus).

ОВЕРЛЕЙНАЯ СХЕМА

Рассмотрим появившийся на Западе современный тип онлайнового научного журнала – оверлейный журнал. Стоимость выпуска оверлейного журнала настолько незначительна, что журнал может позволить себе реализовать схему «бесплатно для автора, бесплатно для читателя».

Оверлейный журнал функционирует в соответствии с декларацией движения Справедливый Открытый доступ – максимально снизить стоимость публикации, опираясь на энтузиазм и ответственность научного сообщества. Оверлейный журнал реализует нетрадиционную схему организации взаимодействия автора и редакции журнала [11, 12], состоящую из следующих шагов:

- автор посылает в журнал статью, предварительно размещенную в архиве в виде препринта,
- журнал проводит рецензирование статьи,
- принятый к публикации вариант статьи получает DOI и вновь размещается в архиве препринтов,

- журнал публикует метаданные статьи и ссылку на полный текст статьи, размещенный в архиве препринтов.

Оверлейную схему журнала можно продемонстрировать на примере онлайнового журнала «Discrete Analysis», реализованного на базе платформы Scholastica (https://scholasticahq.com/) (рис. 4). Платформа предоставляет средства поддержки создания оверлейных журналов, интегрированных с arXiv.

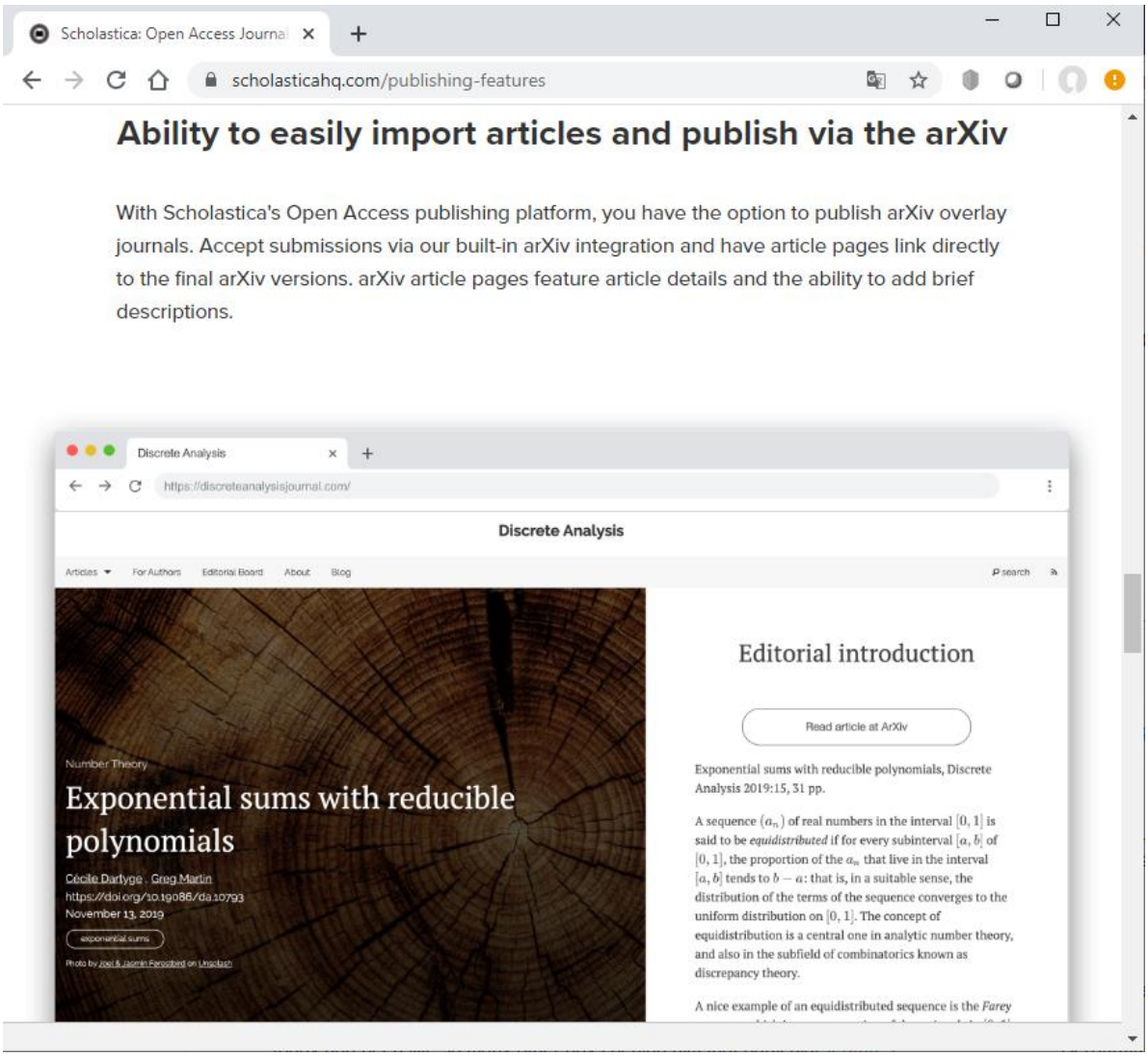


Рис. 4. Платформа Открытого доступа Scholastica (https://scholasticahq.com/) с возможностью создания оверлейного журнала, интегрированного с arXiv

Журнал является рецензируемым, в нем функционирует редакционный совет – как и в традиционном журнале. Журнал принимает статьи, размещенные в arXiv. При поступлении статьи в журнал рецензенты как обычно готовят рецензии,

в которых могут высказать свои замечания по содержанию или оформлению статьи. Автор вносит изменения, предложенные рецензентами и редакторами, и после окончательного одобрения редколлегии статья принимается к публикации. Что происходит далее?

Журнал присваивает статье DOI. Принятая журналом версия статьи снова размещается в arXiv в виде новой версии. В журнале публикуются метаданные статьи (название, авторы, год издания, DOI и др.), а также ссылка на полный текст статьи, размещенный в arXiv.

Приведем еще один пример оверлейного журнала, реализованного на платформе Scholastica. На рис. 5 демонстрируется сайт статьи в журнале «The Open Journal of Astrophysics» [13].

На странице статьи на сайте журнала можно увидеть название текущего раздела журнала, название статьи, ссылки на информационные блоки каждого автора, аннотацию, ссылку на полный текст статьи, размещенный в arXiv, тематические тэги. Тематические теги позволяют показать читателю все статьи журнала, связанные с данной темой. На сайте указано, по какой лицензии доступно использование материалов статьи (все прошедшие рецензирование журнальные статьи публикуются с лицензией CC BY-4.0). Журнал также размещает ссылки на страницы журнала в Twiter и Facebook.

Сервер arXiv предоставляет возможность хранить версии статьи. Как уже было отмечено, после публикации текста первоначального препринта в оверлейном журнале в архиве препринтов появляется новая версия – отредактированная по согласованию с рецензентами и редактором статья. Если автор пожелает дополнить новым материалом текст препринта, то одной из возможных представляется следующая схема работы: автор изменяет текст своего первоначального препринта, а журнальный вариант, получивший DOI, остается без изменений. Но и в текст отрецензированной статьи автор может внести необходимые изменения, связанные, например, с исправлением обнаруженных ошибок. В любом случае у автора появляется возможность развивать свое произведение: исправлять неточности, добавлять новые данные, вносить изменения по результатам обсуждения с коллегами. Если автор развивает свое произведение в течение заметного

периода времени, то статья превращается в «живую публикацию» [14, 15], которую автор поддерживает в актуальном состоянии в ходе своей работы по определенной теме.

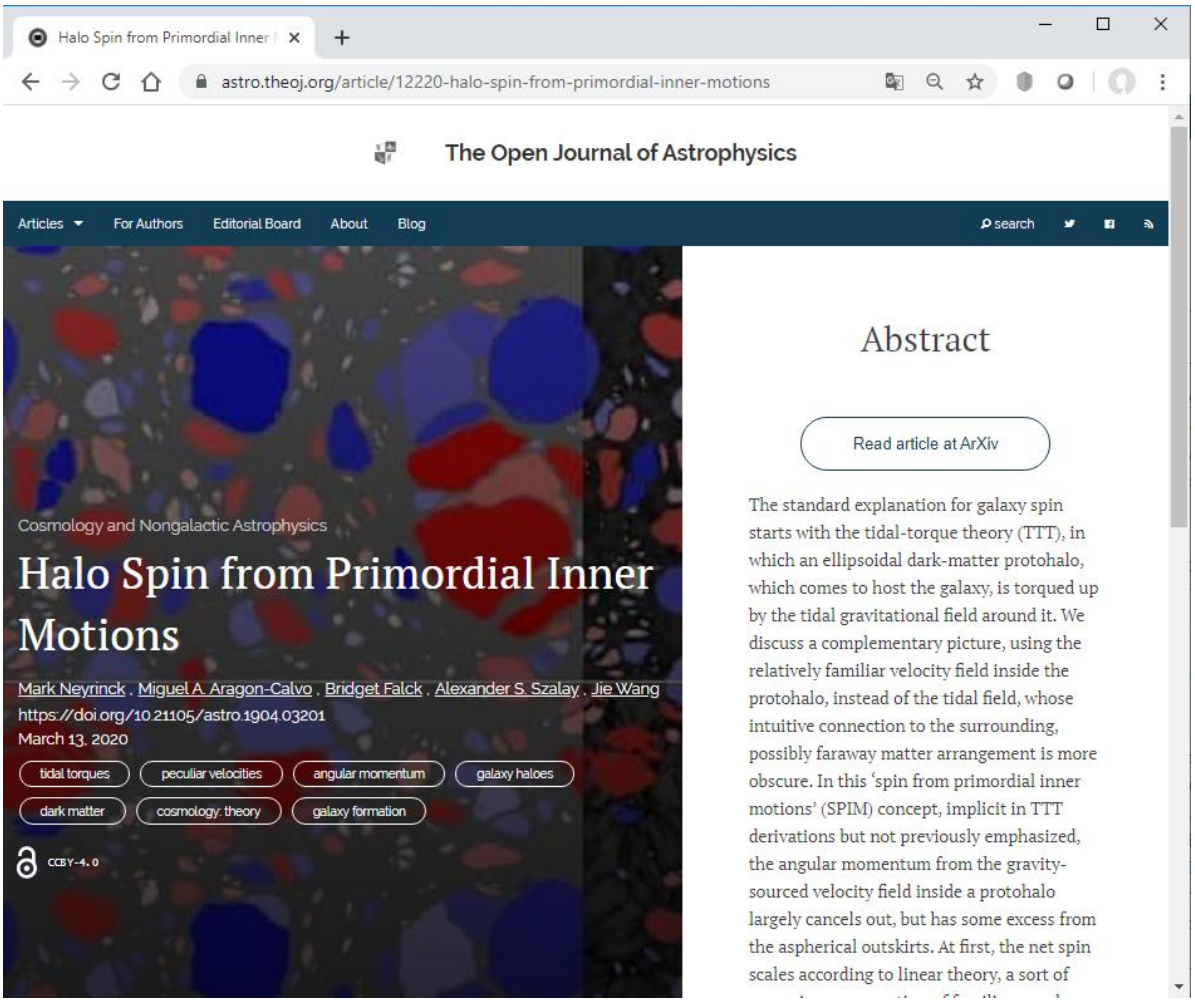


Рис. 5. Статья на сайте журнала «The Open Journal of Astrophysics»

Ничто не мешает автору впоследствии опубликовать свой более полный препринт в другом оверлейном журнале. Это не должно привести к нарушению этических требований. Известно, что большинство издателей не приветствуют повторную публикацию статьи, считая, что повторная публикация может запутать читателя или создать искаженные представления о значимости исследования. Но такие опасения связаны с традиционной схемой публикации статей в журналах, которая ведет к повторным появлениям одного и того же текста в разных журна-

лах. В случае повторной публикации в оверлейном журнале речь идет о появлении версий статьи, видимых в едином месте – на сайте исходного препринта в arXiv. Автор вряд ли будет скрывать историю появления новой версии своей работы: все версии, в том числе опубликованные в журналах, видны на сайте статьи в arXiv. Публикация статьи в нескольких оверлейных журналах будет отражать более высокую оценку статьи, формируемую не на основе подсчета цитирований, а на основе полученных положительных рецензий.

Важно подчеркнуть, что оверлейная схема журнала реализована не в единичном проекте. Еще одним из примеров поддержки создания оверлейных журналов является платформа Episciences [16]. В настоящее время на этой платформе размещено полтора десятка журналов Открытого доступа, реализующих оверлейную схему. Организаторы Episciences призывают другие журналы присоединиться к сообществу.

В чем состоят преимущества оверлейного журнала?

Прежде всего, снижаются накладные производственные расходы: оверлейный журнал занимается своей основной функцией – проведением научного рецензирования статей, а вопросы организации хранения полных текстов статей решаются на уровне архива препринтов. Архив препринтов на своей базе может обслуживать одновременно несколько оверлейных журналов.

На наш взгляд, главное достоинство оверлейной схемы состоит в том, что автор не теряет связь с опубликованным текстом. Вспомним, что изменить текст в опубликованной журнальной статье довольно трудно и в случае печатной версии, и в случае онлайновой версии журнала: требуются обсуждения и согласования в «ручном» режиме. В случае, когда полные тексты препринта и вариантов статей находятся в архиве препринтов, автор может самостоятельно вносить нужные изменения. Вероятно, в разных архивах препринтов существуют разные политики по отношению к процедуре замены ранее загруженных текстов, но есть основания предполагать, что архив препринтов более лояльно относится к внесению изменений и созданию версий, чем традиционный журнал.

Развивающаяся инфраструктура архивов Открытого доступа и серверов препринтов служит надежной основой для появления оверлейных журналов, предлагающих авторам и читателям бесплатные услуги. К сожалению, в России мы не

наблюдаем заметного технологического развития архивов препринтов. В работе [17] в 2009 г. отмечалось, что наиболее естественным местом опубликования результатов научных исследования является сайт организации, где проводились исследования. Но за прошедшее десятилетие в российской инфраструктуре научных публикаций не произошло каких-либо заметных сдвигов: научные организации и вузы по-прежнему недооценивают свою важную миссию – создание архивов для публикации на своих сайтах научных результатов в виде препринтов. Можно утверждать, что в нашей стране культура издания препринтов пока находится в зачаточном состоянии. Поэтому о массовом появлении полновесных оверлейных журналов в России пока говорить рано.

ЗАКЛЮЧЕНИЕ

Оверлейный онлайновый журнал опирается на общедоступные серверы препринтов, которые поддерживаются научными учреждениями и вузами. Автор представляет в журнал свою статью (препринт) непосредственно с сервера препринтов. Оверлейный журнал проводит рецензирование поступившей статьи. В случае принятия статьи к публикации журнал размещает на своем сайте метаданные статьи, а сама статья (ее скорректированный полный текст) вновь размещается на сервере препринтов. Сервер препринтов дает возможность автору не терять связь со своей работой: автор может продолжать развивать свою статью и создавать новые более полные версии. Оверлейная схема работы журнала не перегружает функциональность сервера препринтов, но ведет к снижению расходов издателя и способствует реализации принципа «бесплатно для автора, бесплатно для читателя».

Оверлейный онлайновый журнал – это не только удачная организационная схема. Оверлейный журнал имеет существенно более важное влияние на издательский процесс: значительное снижение стоимости производства научного журнала позволяет возвратить публикационную деятельность под контроль научного сообщества.

Благодарности

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект 19-01-00069-а.

СПИСОК ЛИТЕРАТУРЫ

1. *Горбунов-Посадов М.М.* Интернет-активность как обязанность ученого. URL: https://keldysh.ru/gorbunov/duty.htm (редакция от 16.05.2020).

2. *Peter Coles.* The Age of Preprints. Post in «The Open Journal of Astrophysics». 23.09.2020. URL: https://astro.theoj.org/post/674-the-age-of-preprints

3. *Горбунов-Посадов М.М.* Препринты ИПМ им. М.В. Келдыша. URL: http://www.keldysh.ru/gorbunov/preprints.htm (редакция от 06.04.2019).

4. *Полилова Т.А.* Научная публикация в России: интеллектуальные права // Препринты ИПМ им. М.В. Келдыша. 2019. № 56. С. 1–4. http://doi.org/10.20948/prepr-2019-56. URL: http://keldysh.ru/papers/2019/prep2019_56.pdf

5. *Polilova T.A.* Ethical norms and legal framework of scientific publication // Mathematica Montisnigri, V. XLV (2019). P. 129–136. http://doi.org/10.20948/mathmontis-2019-45-11 URL: http://www.montis.pmf.ac.me/vol45/11.pdf.

6. *Полилова Т.А.* О лицензионном договоре на издание служебного произведения // Электронные библиотеки. 2019. Т. 22, № 2. С. 119–141. http://doi.org/10.26907/1562-5419-2019-22-2-119-141; URL: http://ojs.kpfu.ru/index.php/elbib/article/view/973

7. *Беляева Светлана.* Цена открытости: Во что обойдется переход к Open Access? // Поиск. 24.05.2019. URL: https://www.poisknews.ru/skript/czena-otkrytosti-vo-chto-obojdetsya-perehod-k-open-access/

8. *Adler J.R., Chan T.M., Blain J.B., Thoma B., Atkinson P.* OpenAccess: Free online, open-access crowdsource-reviewed publishing is the future; traditional peer-reviewed journals are on the way out // Canadian Journal of Emergency Medicine. 2019. 21(1). P. 11–14. URL: https://doi.org/10.1017/cem.2018.481

9. *Fair Open Access Alliance.* URL: https://www.fairopenaccess.org/

10. *Free Journal Network.* URL: https://freejournals.org/

11. *Wikipedia. Overlay journal.* URL: https://en.wikipedia.org/wiki/Overlay_journal

12. *Herman E., Akeroyd J., Bequet G., Nicholas D., Watkinson A.* The changed – and changing – landscape of serials publishing: Review of the literature on emerging models. First published: 17 February 2020. https://doi.org/10.1002/leap.1288
URL: https://onlinelibrary.wiley.com/doi/full/10.1002/leap.1288

13. *The Open Journal of Astrophysics*. URL: https://astro.theoj.org/

14. *Горбунов-Посадов М.М.* Живая публикация.
URL: https://keldysh.ru/gorbunov/live.htm

15. *Gorbunov-Posadov M.M., Polilova T.A.* Tools to Support Scientific Online Publishing // Programming and Computer Software. May 2019. V. 45, Issue 3. P. 116–120. URL: https://link.springer.com/article/10.1134%2FS0361768819030046

16. *Episciences. Overlay Journal Platform.* URL: www.episciences.org/?lang=en

17. *Полилова Т.А.* Инфраструктура научных публикаций // Препринты ИПМ им. М.В. Келдыша. 2009. № 15. 30 с.
URL: http://library.keldysh.ru/preprint.asp?id =2009-15

PREPRINT AS THE MATERIAL FOR AN OVERLAY JOURNAL

T. A. Polilova[0000-0003-4628-3205]

Keldysh Institute of Applied Mathematics Russian Academy of Sciences

polilova@keldysh.ru

Abstract

The Open access movement has a long history. In 2002 the Budapest Open access initiative was first announced. However, the problem of Open access has not yet been fully and definitively resolved. In 2018 The European Union has adopted Plan S, which calls for making Open access a reality by 2020. Plan S emphasizes the importance of self-archiving of articles and the role of Preprint's archives (servers) for scientific results placement. It is noted that Preprint archives have a great potential for editorial and publishing innovations. Scientific journals with limited reader access that operate on a commercial basis do not give up their positions. But even here we see some progress. Journals have become less rigid in their policy towards preprints and post-prints.

More and more foreign scientists are becoming adherents of the "Fair open access" movement, which offers a new organizational solution. The journal must have a scientific organization or non-profit Foundation as a founder, that hires a group of executors to provide editorial and publishing services. Editors and publishers should not have their own commercial interests. The scientific journal should be funded from the general contribution of organizations.

The article considers a modern type of online scientific journal — the overlay journal. The cost of an issue of the overlay journal is so low that the journal can easily implements the "free for the author, free for the reader" scheme. The overlay journal is based on the public servers of preprints. The online overlay journal reviews the article received from the archive. If the article is accepted for publication, the article metadata is published on the journal website, and the full text of corrected article is re-archived. This way of working does not overload the archive functionality, but it allows to reduce the financial burden on the overlay journal.

Keywords*: scientific journal, Fair open access, Open archive, server of preprints, overlay journal.*

REFERENCES

1. *Gorbunov-Posadov M.M.* Internet-aktivnost kak obiazannost uchenogo. URL: https://keldysh.ru/gorbunov/duty.htm (redaktsiia ot 16.05.2020).

2. *Peter Coles.* The Age of Preprints. Post in «The Open Journal of Astrophysics». 23.09.2020. URL: https://astro.theoj.org/post/674-the-age-of-preprints

3. *Gorbunov-Posadov M.M.* Preprinty IPM im. M.V. Keldysha. URL: http://www.keldysh.ru/ gorbunov/preprints.htm (redaktsiia ot 06.04.2019).

4. *Polilova T.A.* Nauchnaia publikatsiia v Rossii: intellektualnye prava // Preprinty IPM im. M.V. Keldysha. 2019. № 56. S. 1–24. https://doi.org/10.20948/prepr-2019-56, URL: http://keldysh.ru/papers/2019/prep2019_56.pdf

5. *Polilova T.A.* Ethical norms and legal framework of scientific publication // Mathematica Montisnigri. 2019. Vol. XLV. P. 129–136. https://doi.org/10.20948/math-montis-2019-45-11, URL: http://www.montis.pmf.ac.me/vol45/11.pdf

6. *Polilova T.A.* O litsenzionnom dogovore na izdanie sluzhebnogo proizvedeniia // Elektronnye biblioteki. 2019. T. 22, № 2. S. 119–141. http://doi.org/10.26907/1562-5419-2019-22-2-119-141;
URL: http://ojs.kpfu.ru/index.php/elbib/article/view/973

7. *Beliaeva Svetlana*. Tsena otkrytosti: Vo chto oboidetsia perekhod k Open Access? // Poisk. 24.05.2019.
URL: https://www.poisknews.ru/skript/czena-otkrytosti-vo-chto-obojdetsya-perehod-k-open-access/

8. *Adler J.R., Chan T.M., Blain J.B., Thoma B., Atkinson P.* OpenAccess: Free online, open-access crowdsource-reviewed publishing is the future; traditional peer-reviewed journals are on the way out // Canadian Journal of Emergency Medicine. 2019. 21(1). P. 11–14. URL: https://doi.org/10.1017/cem.2018.481

9. *Fair Open Access Alliance*. URL: https://www.fairopenaccess.org/

10. *Free Journal Network.* URL: https://freejournals.org/

11. *Wikipedia. Overlay journal*.
URL: https://en.wikipedia.org/wiki/ Overlay_journal

12. *Herman E., Akeroyd J., Bequet G., Nicholas D., Watkinson A.* The changed – and changing – landscape of serials publishing: Review of the literature on emerging models. First published: 17 February 2020. https://doi.org/10.1002/leap.1288, URL: https://onlinelibrary.wiley.com/doi/full/10.1002/leap.1288

13. *The Open Journal of Astrophysics*. URL: https://astro.theoj.org/

14. *Gorbunov-Posadov M.M.* Zhivaia publikatsiia.
URL: https://keldysh.ru/ gorbunov/live.htm

15. *Gorbunov-Posadov M.M., Polilova T.A.* Tools to Support Scientific Online Publishing // Programming and Computer Software. May 2019. V. 45, Issue 3. P. 116–120. URL: https://link.springer.com/article/10.1134%2FS0361768819030046

16. *Episciences. Overlay Journal Platform*. URL: www.episciences.org/?lang=en

17. *Polilova T.A.* Infrastruktura nauchnykh publikatsii // Preprinty IPM im. M.V. Keldysha. 2009. № 15. 30 s.
URL: http://library.keldysh.ru/preprint.asp?id=2009-15

СВЕДЕНИЯ ОБ АВТОРЕ

ПОЛИЛОВА Татьяна Алексеевна – старший научный сотрудник Института прикладной математики им. М.В. Келдыша РАН, доктор физико-математических наук, лауреат Премии Президента РФ в области образования;

Tatyana Alekseevna POLILOVA – senior researcher of the Institute of Keldysh Institute of Applied Mathematics Russian Academy of Sciences.

email: polilova@keldysh.ru

Материал поступил в редакцию 19 ноября 2020 года

УДК 004.414.3

РАСКАДРОВКА КАК ОДНО ИЗ ПРЕДСТАВЛЕНИЙ СЦЕНАРНОГО ПРОТОТИПА КОМПЬЮТЕРНЫХ ИГР

Г. Ф. Сахибгареева[1, [0000-0003-4673-3253]], О. А. Бедрин[2, [0000-0003-3300-4318]], В. В. Кугуракова[3, [0000-0002-1552-4910]]

[1,2,3]*Институт информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета*

[1]gulnara.sahibgareeva42@gmail.com, [2]simplavero@gmail.com, [3]vlada.kugurakova@gmail.com

Аннотация

Работа посвящена изучению и усовершенствованию процесса проектирования, разработки и тестирования повествования видеоигр. Изучены существующие практики написания и поддержки в актуальном состоянии сценария интерактивных произведений. Сформулированы определение сценарного прототипа, а также требования к его форме. Выдвинута идея об эффективности автоматизации создания сценарного прототипа в виде инструмента-генератора. Составлено видение такого инструмента. Представлено влияние такого инструмента на порядок разработки. Реализован компонент инструмента и проведен эксперимент, который доказывает эффективность на таком примере, как генерация раскадровки из текста. Сформулированы планы на будущую разработку.

***Ключевые слова**: компьютерные игры, разработка видеоигр, интерактивное повествование, сценарный прототип, нарративный дизайн, сценаристика, игровая документация, раскадровка, генерация раскадровки, интерактивная раскадровка.*

ВВЕДЕНИЕ

Разработка видеоигр – длительный и дорогостоящий процесс, поэтому любые оптимизация и автоматизация могут сыграть ключевую роль в успехе финального продукта.

Сложность разработки видеоигр заключается прежде всего в их составе. Они включают в себя визуализацию, аудиоряд, интерактивное взаимодействие,

правила игры и историю. Каждый компонент тесно связан с остальными, и все вместе они дают результат, уникальный для любого медиа, – опыт интерактивного потребления контента с драматургической подоплекой. Иными словами – игроки сами творят историю.

Индустрия разработки игр – молодое и бурно развивающееся направление. Конкретно повествовательная составляющая долго оставалась скорее бременем, чем мощным инструментов. Только в последние годы в данном направлении начинают появляться своя теория и свои эксперты.

В рамках данной работы проведен анализ существующих практик разработки интерактивного повествования, а также результатов попыток автоматизации этого процесса.

Глобальная цель этой работы – познакомить с процессом разработки инструмента, который призван автоматизировать генерацию прототипа повествования видеоигр, а также продолжить работу по исследованию особенностей процесса разработки видеоигр.

В разделе 1 «Инструменты сценариста видеоигр» приведен анализ ряда форматов документации, практик разработки, а также программного обеспечения, которое применяется в разработке видеоигр.

В разделе 2 «Сценарный прототип» даны определение сценарного прототипа, обоснование этого определения, а также сравнительный анализ этой практики с некоторыми инструментами, приведенными в разделе 1.

В разделе 3 «Генерация сценарного прототипа» дано обоснование необходимости и пользы автоматизации процесса создания сценарного прототипа в форме генератора.

В разделе 4 «Существующие решения визуализации» проанализированы существующие инструменты визуализации, которые доказывают, что генерация визуализации из текста возможна.

В разделе 5 «Видение инструмента генерации сценарного прототипа» представлены схема и видение генератора сценарных прототипов, а также приведен пример внедрения этого инструмента в конвейер разработки повествования.

В разделе 6 «Прогресс в разработке инструмента» представлен прогресс в разработке генератора.

В разделе 7 «Видение процесса генерации событий в пространстве» сформулировано видение процесса генерации визуализации для сценарного прототипа для модуля визуализации.

В разделе 8 «Реализация процесса генерации событий в пространстве» приведена техническая реализация.

В разделе 9 «Результаты работы процесса визуализации» приведены пример генерации раскадровки, а также результаты эксперимента.

В разделе 10 «Дальнейшее развитие» сформулированы планы на будущее.

1. ИНСТРУМЕНТЫ СЦЕНАРИСТА ВИДЕОИГР

Для налаживания коммуникаций с командой разработки и донесения основных идей и важных нюансов сценаристы используют различную документацию. Некоторые технические сведения используются только в процессе разработки; другие содержат текст, который игроки увидят в видеоигре в качестве контентной составляющей. Каждый формат документации призван решить специфические задачи реализации повествовательной составляющей игры. Разработка не может идти с применением единственного формата, т. к. каждый из них имеет свои преимущества и недостатки.

В настоящей раздел представлен анализ таких форматов, как текстовый документ, wiki-разметка, таблица, раскадровка, а также такой практики, как бумажный прототип. Помимо этого, проанализировано программное обеспечение, которое может применяться в работе игровых сценаристов: Twine [1], RenPy [2], Articy Draft [3].

Перечень проанализированных инструментов неисчерпывающий. Помимо них сценарист может использовать в своей работе проприетарные инструменты игровой компании, такие как модули редактирования диалогов и квестов, реализованные в игровых движках Unity [4] или Unreal Engine [5], или другие инструменты и сервисы, которые адаптируются для задач разработки, например, Miro [6], diagrams.net [7] и другие. Анализ проведен нами с точки зрения эффективности в процессе разработки игрового повествования.

1.1. Форматы документации

Текстовый документ

Явление интерактивности скрывает потенциал, который не раскрыт в полной мере. Это значит, что необходимы специализированные форматы и инструменты для работы с формами медиа, которые содержат в себе интерактивность.

Пока индустрии разработки видеоигр не хватает собственной фундаментальной теории и глубокого осознания специфики интерактивного повествования видеоигр, поэтому применяются практики и инструменты других медиа. Например, в случае видеоигр уместны драматургические приемы. Когда речь идет о линейном повествовании или небольшом количестве выбора или о кат-сценах, видеороликах, которые переключают с активного игрового процесса на пассивный просмотр кинематографических отрывков, действуют те же правила погружения, как и для зрителя. Различие состоит в том, что такие вставки прерывают процесс игры, что иногда бывает неуместно.

Другой пример – голливудский стандарт написания документации. Он общепринят во всем мире и применяется для форматирования сценариев видеоигр.

Однако, когда речь заходит о поддержании объемного разветвленного повествования, формата текстового документа оказывается недостаточно. Для создания истории с возможностью выбора необходимы гибкие инструменты, которые позволяют создавать, редактировать и вести учет каждого разветвления в структуре игрового повествования. Редактирование текста – это трудоемкий процесс, которые лишен возможности отслеживания комплекса зависимостей событий друг от друга.

Эффективность восприятия текста сравнительно низкая, в отличие от визуальных и интерактивных произведений. Было доказано, что эффективность визуализации [8] и, тем более, погружения [9] значительно выше, чем у передачи информации через текст. Таким образом, текст, сопровождаемый визуализацией или визуализированный в полном объеме, воспринимается лучше, а это значит, что внутри команды разработки увеличивается вероятность взаимного понимания сюжета разрабатываемой игры.

Wiki-разметка

Формат wiki-разметки наследует недостатки текста – это отсутствие гибкости, сложность в поддержании актуальности, а также необходимость в длительном и вдумчивом изучении. Однако он имеет и преимущества: формат wiki-разметки – это возможность создания веб-энциклопедии о внутриигровом мире. В каждой тематической статье возможно размещение ссылок на другие статьи, что создает возможность нелинейного изучения информации.

Данный формат применим в работе сценаристов для специфической цели хранения всеобщего свода правил и истин, касающихся игрового мира. Кроме этого опубликованные wiki-ресурсы позволяет создать некоторую трансмедийность, в которой игроки могут познакомиться с миром игры как в рамках игрового процесса, так и вне игрового пространства. Однако чаще такие ресурсы создаются усилиями как раз фан-сообществ, что еще раз доказывает, что данный формат имеет успех в представлении больших объемов информации в виде, привлекательном для изучения.

Таблица

Таблица позволяет представлять информацию в двумерном и даже в N-мерном пространстве, в котором каждая ячейка определяется правилами и связью с другими ячейками. Рассмотрим преимущества и недостатки табличного представления на примере (см. рис. 1).

№	Название	Персонаж	Реплика персонажа	1 этап	2 этап	3 этап	4 этап	5 этап
1	Именинный торт	Инна	“Сегодня у Оли День рождения. Давай испечем ей торт.”	Купить продукты	Выбрать рецепт	Испечь коржи	Собрать торт	Подарить торт

Рис. 1. Пример квеста в формате таблицы

В примере приведен фрагмент воображаемого мобильного квеста. Для каждого внутриигрового задания выделены следующие параметры: идентификационный номер, название, имя персонажа, который дает игроку задание, его реплика для дачи задания и этапы прохождения задания. Рассмотрим цель каждого параметра:

- идентификационный номер необходим для ускорения коммуникаций в команде разработки;
- название задания также необходимо для коммуникации, но помимо этого оно отображается в самой игре;
- игрок должен каким-то образом получить задание, в данном случае его выдает конкретный персонаж, который с этой целью произносит определенную реплику;
- для выполнения задания игрок должен пройти конечное число этапов.

Табличное решение можно расширять по количеству параметров и строк до объема, необходимого в разработке. Табличное решение эффективнее текстового, т. к. наглядно представляет информацию в структурированном виде. Заметить в таком представлении ошибки и исправить их значительно проще, в особенности, если в таблице настроены формулы и скрипты.

Таблицы подходят не только для квестов. В таком формате можно хранить информацию о необходимости визуализации и аудиального сопровождения, указать специальные требования. Кроме этого возможно прописать формулы для составления текстов из внутриигровых сущностей.

Таблицы – хороший технический инструмент, призванный структурировать информацию и организовать работу команды. Однако он не подходит для презентации истории игры и последующего тестирования и оценки, т. к. таблица не имеет ничего общего с игровым процессом.

1.1. Раскадровка

Раскадровка используется в производстве фильмов, рекламы и других аудиовизуальных произведений. Это последовательность кадров, которая фиксируют в себе основные сцены и кадры из будущего произведения. Степень детализации при этом может быть разной.

Если добавить для каждого кадра возможность выбора, то следующий кадр, который станет доступен для просмотра, будет определяться эти выбором. Получается интерактивная раскадровка. Данный термин не является общепринятым и используется в контексте данной статьи как авторское определение.

Прямой аналогией для интерактивной раскадровки является жанр визуальных новелл. В них есть фоновая картинка, картинки персонажей, окна с именами и репликами персонажей, а также окно выбора реплики или выбора дальнейшего развития событий.

Как можно оценить визуализацию классической раскадровки, так можно протестировать интерактивную раскадровку на различные параметры. Можно оценить наличие персонажей в той или иной сцене, порционность текста в диалогах и записках, уместность аудиального сопровождения и цветового решения.

Раскадровка, как способ визуализации, привлекательнее и эффективнее, чем текст или таблица. Этот упрощенный формат поддается тестированию и сбору оценок от целевой аудитории. Данный формат также примечателен тем, что реализует интерактивную природу игры.

Минусом раскадровки является статичность картинки. Основная доля видеоигр – это произведения, в которых игрок не просто видит набор кадров, он самостоятельно перемещается в пространстве и выбирает способ взаимодействия с доступными внутриигровыми объектами.

1.2. Бумажный прототип

Бумажный прототип – это гибкий инструмент разработки и тестирования игрового повествования. Его отличие в том, что он позволяет вносить изменения в историю на ходу, во время проверки на целевой аудитории.

Для создания бумажного прототипа необходимо подобрать необходимые материалы: а) поле для развития действий, карту, разлинованный ватман или шахматную доску; б) фигуры, которые будут олицетворять персонажей игроков и игровых персонажей; в) кубики или приложение с генерацией чисел для создания случайности в развитии событий и г) специфические элементы, присущие проектируемому повествованию. Если сценарист способен провести игроков через всю историю с помощью подручных средств и достаточно увлечь участников, то есть вероятность, что история будет также увлекательна и в видеоигре.

Однако здесь выступает другая особенность видеоигр – наличие мануального взаимодействия, а также восприятие аудиальной и визуальной составляющих через устройства ввода/вывода компьютера, телефона и других платформ (см., например, [10, 11]). Такая специфика, как целевое устройство и атмосфера, в

которых находится игрок, важна с точки зрения моделирования опыта. Бумажный прототип не дает возможности воссоздания реалистичных условий процесса игры. С помощью бумажного прототипа можно проверить специфические характеристики интерактивного повествования, такие как погружение, темпоритм или эмпатия.

Погружение [12] – состояние, в котором игрок забывает о времени и реальности, проникается историей, осознает ее важность для себя. Погружение – это знакомое состояние, в котором каждый может обнаружить себя, когда читает книги и смотрит фильмы. Это чувство оторванности от происходящего. Тот же эффект уместен для интерактивного повествования.

Темпоритм [13] – частота и плотность подачи истории. Игрок знакомится с миром игры постепенно. Кроме этого должна сохраняться интрига, которая удержит игрока в процессе изучения истории. Темпоритм регулируется исключительно с помощью тестирования, и бумажный прототип позволяет это сделать.

Эмпатия [14] – это важное чувство, которое рождается у игрока по отношению к персонажу, за которого он играет, а также по отношению к другим персонажам. Каждый находит в героях истории что-то своё. Но бывают неудачные персонажи, которые либо сами по себе плохо продуманы, либо не соответствуют ожиданиям целевой аудитории, текущему времени. Данный параметр также проверяется только на практике в ходе тестирования.

Таким образом, бумажный прототип является эффективным и дешевым инструментов тестирования игровой истории на ранних стадиях [15]. Однако еще один минус бумажного прототипа состоит в том, что каждый раз при необходимости провести тестирование необходимо организовать встречу и задействовать сценариста.

1.3. Программное обеспечение

Свойства, важные для инструментария, который применяет игровой сценарист, – это гибкость, возможность поддержки разветвленной системы и ее быстрого редактирования. Для этого сценаристы применяют дополнительное специализированное программное обеспечение или адаптируют существующие сер-

висы под свои задачи. Ниже представлен анализ некоторых инструментов, которые принципиально отличаются друг от друга функционалом и возможностями в разработке: Twine, RenPy, Articy Draft.

Twine

Twine – это open-source движок для создания текстовых игр. Инструмент снижает порог вхождения в игровую индустрию благодаря трем ключевым преимуществам:

- построение историй происходит в виде системы карточек, навыки создания html-разметки при этом необходимы минимально;
- доступны настройки стиля текста, макросов, добавление условий, переменных, картинок и звуков;
- доступен экспорт проекта в виде html-файла; его можно опубликовать в виде веб-страницы.

Результат работы в Twine можно экспортировать в Unity с помощью ассета Crandle [16].

RenPy

RenPy – бесплатный open-source движок для разработки визуальных новелл. Три ключевых преимущества RenPy:

- редактор кода находится в самом движке;
- используется достаточно легкий в изучении язык программирования Python;
- имеется встроенный гайд для быстрого освоения инструмента.

В проект в RenPy можно добавлять стандартные компоненты новелл – текст, фоновые изображения, персонажей, музыку, звуки, анимации. Визуальная новелла эффективнее текстовой игры, т. к. включает в себя визуализацию.

Articy:draft

Articy:draft – это приложение для визуализации и организации игровых компонентов. Разработчики относят его к приложениям для разработки point-and-click адвенчур [17], но оно применимо при разработке игр других жанров.

Articy:draft позволяет систематизировать информацию, которая необходима в течение разработки игры в стандартной форме. Систему хранения сущностей в Articy:draft можно сравнить с диаграммой связей или с системой wiki-разметки.

С точки зрения разработки сценарного прототипа у Articy:draft есть следующие три преимущества:

- экспорт данных доступен в формате json, что делает инструмент универсальным;
- в Articy:draft можно хранить разнообразную информацию: персонажей, диалоги, квесты, инвентарь, достижения и других игровые сущности;
- принцип визуального программирования снижает порог вхождения для начинающих специалистов.

Минус Articy:draft – в его стоимости: система доступна для крупных компаний. В случае с независимыми разработчиками использование Articy:draft может оказаться неоправданным. Функционал Articy:draft позволяет назвать его универсальным инструментом благодаря встроенным шаблонам и готовым элементам пользовательского интерфейса. Articy:draft позволяет разработать быстрые прототипы для системы диалогов и узкого перечня жанров.

2. СЦЕНАРНЫЙ ПРОТОТИП

Сценарный прототип – это интерактивный цифровой прототип повествовательной составляющей видеоигры. Приведем обоснование определения.

а) *Цифровой формат* более точно моделирует игровой процесс будущей игры. Игрок будет проводить время за компьютером или телефоном, поэтому важно соблюсти это физическое взаимодействие на ранних этапах, чтобы не исказить впечатления и получить релевантные отзывы о прототипе.

б) *Повествовательная компонента* неотделима от игрового процесса. В рамках определения допускается, что прототип фокусируется на повествовании и игнорирует интерактивность ровно до тех пор, пока это не принципиально для повествования. В случае механик боя или для создания неожиданности события возможно применение генерации случайного исхода, чтобы смоделировать эмо-

ции от игрового процесса. В случае же повествовательных механик подразумевается, что они будут реализованы в сценарном прототипе. Когда речь идет о диалоге или изучении текстовых носителей, будь то записок или энциклопедий, их также следует включить в прототип. Выбор зависит от конкретного проекта, а также от целей прототипирования.

Иными словами, сценарный прототип – это черновая разработка, соответствующая по жанру разрабатываемой игре; она иметь “заглушки” на геймплей и игровые механики; разработана на целевом движке.

Соответствие сценарного прототипа жанру, а также его разработка на целевом движке – рекомендации, устраняющие возможные трудности с имплементацией сценарного прототипа в игру. К тому же, всегда проще нарастить функционал сценарного прототипа, чем разрабатывать игру с нуля. Иметь готовый сценарный прототип для каких-то игр означает иметь одну из итераций разработки. Таким образом, сценарный прототип сокращает объём работ в течение этапа разработки.

Сценарный прототип – это попытка на время изолировать историю от игры, чтобы протестировать её на такие характерные параметры интерактивного повествования, как погружение, темпоритм и эмпатию. Это решение, позволяющее сэкономить проектные ресурсы, протестировать историю до начала разработки контентной составляющей игры. Он применим как для небольших, так и для крупных проектов, как для разработки с нуля, так и для доработки дополнительных блоков существующей игры.

Разработка сценарного прототипа на основе работы сценариста, представленной в виде технической документации, – работа нарративного дизайнера[1].

В качестве обоснования актуальности данного подхода, приведем сравнительный анализ с другими форматами и практиками разработки (см. рис. 2). Критерии оценки для сравнительного анализа следующие: визуализация, время на изучение, поддержка разветвления, возможность правок, тестирование игрового опыта. Сравниваются инструменты друг относительно друга. Оценка зависит от названия колонки и означает объем возможности или количество необходимых

[1] Нарративный дизайнер – специализация гейм-дизайнера; дизайнер по имплементации истории в игру.

ресурсов: мало, средне, много. Цветом выделена степень эффективности: зеленый – отлично подходит для решения поставленных задач, оранжевый – подход имеет недостатки, красный – формат плохо подходит для работы.

Главная задача сравнения – показать, что сценарный прототип относится к эффективным практикам проектирования, разработки и тестирования интерактивного повествования.

№	Артефакт разработки	Визуализация	Время на изучение	Поддержка разветвления	Возможность правок	Тестирование игрового опыта
1	Текстовый документ	мало	много	мало	мало	мало
2	Wiki-разметка	средне	много	средне	мало	мало
3	Таблица	средне	средне	средне	средне	мало
4	Раскадровка	много	мало	мало	мало	мало
5	Бумажный прототип	средне	мало	много	много	много
6	Сценарный прототип	много	мало	много	много	много

Рис. 2. Сравнение инструментов сценариста

По результатам анализа можно пройти к следующим выводам:

- текстовый документ почти не поддерживает визуализацию, требует много времени на изучение, почти не поддерживает разветвление сюжета, его трудно править, и он совершенно не подходит для тестирования игрового опыта;
- wiki-разметка, как частный случай текстового документа, наследует большую часть проблем этого формата, однако поддерживает нелинейность в изучении текста, а также может содержать в себе больше визуализации;
- таблицы в целом применимы для разработки разветвленного повествования, но не подходят для его тестирования;
- раскадровка хорошо визуализирует историю и сокращает время изучения, но кроме этого больше никакими преимуществами не обладает; кроме

этого раскадровка бессмысленна без сопровождающего текстового документа, комментариев или устного объяснения художника и сценариста;

- бумажный прототип – принципиально иной уровень разработки и тестирования интерактивного повествования, однако визуализация почти отсутствует из-за ограничений данного формата;
- сценарный прототип – инструмент, который во многом подходит для разработки, проектирования и тестирования интерактивного повествования.

3. ГЕНЕРАЦИЯ СЦЕНАРНОГО ПРОТОТИПА

Задача сценариста – создать историю, которая найдет себя в игровом процессе. Задача нарративного дизайнера – внедрить историю в игру или игру в историю так, чтобы они оставались неразрывным целым на протяжении всего игрового процесса. Нарративный дизайнер ближе к непосредственной разработке, в отличие от сценариста, поэтому наличие навыков программирования для них – это значительное преимущество, а для сценариста это скорее лишняя нагрузка.

Создание сценарного прототипа подразумевает работу в игровом движке и иногда включает в себя программирование. По сути это прототипирование компонентов игры, а затем проработка его до финальной версии.

В сфере информационных технологий прототипирование – неотъемлемая часть разработки крупных и небольших проектов. Однако большая часть деятельности в данной сфере посвящена автоматизации. Автоматизация прототипирования – это то, что в перспективе сэкономит время, деньги и другие ресурсы на проекте. Это уместно и для видеоигр.

Генерация сценарного прототипа вместо ручной сборки поможет решить две задачи: 1) разгрузить нарративного дизайнера и 2) снизить порог вхождения нарративных дизайнеров в индустрию. Кроме этого оптимизация нагрузки на специалистов высвободит как финансовые, так и кадровые ресурсы, которые могут быть перераспределены в проекте более эффективным образом.

Первая задача связана с тем, что на предварительном этапе разработки нарративный дизайнер занимается не только тем, что собирает сценарный прототип. Кроме этого ему необходимо создать концепцию внедрения повествования в игру. Если снять с нарративного дизайнера одну задачу, он сможет сконцентрироваться на другой.

Вторая задача также частично связана с разгрузкой других специалистов. Например, если тот же сценарный прототип будет сгенерирован, нарративный дизайнер сможет самостоятельно находить какие-то нестыковки в истории и исправлять их раньше. Тем более при этом не будут задействованы такие специалисты, как художники, программисты и разработчики дорогостоящей трехмерной графики.

Таким образом сформулирована ценность инструмента: генератор сценарных прототипов может сэкономить все возможные ресурсы, которые есть на проекте по разработке игр. Более того, инструмент применим не только в игровой индустрии, но и в индустрии разработки serious games, тренажеров и симуляторов.

Основным сценарием использования генератора сценарного прототипа является следующий план действий: 1) ввести в инструмент текстовые данные, которые содержат в себе как наработки сценариста, так и наработки смежных специалистов, в том числе существующие наработки художника; 2) генератор анализирует и достраивает видение игры и формирует сценарный прототип, а также статистическую информацию и визуализацию структур повествования; 3) пользователь получает результаты генерации, скачивает их и отправляет на тестирование; при необходимости вносит правки.

Более подробно видение инструмента представлено в разделе 5.

4. СУЩЕСТВУЮЩИЕ РЕШЕНИЯ ВИЗУАЛИЗАЦИИ

Сценарный прототип – комплексный продукт, как и видеоигры, которые могут включать в себя текст, визуализацию, звук, интерактивное взаимодействие и возможность выбора.

В рамках данной работы был проанализирован ряд работ, которые так или иначе затрагивают тему эффективной, быстрой и функциональной визуализации как в автоматизированной, так и в ручном форме.

Инструменты разделены на генераторы визуализации из текста и другие инструменты, которые повлияли на видение инструмента генерации сценарного прототипа видеоигр.

4.1. Генерация визуализации из текста

Существует множество решений для генерации визуального контента на основе текста на естественном языке, которые отличаются целями и подходами. Так, например, в работе 2016 года [18] рассматриваются 26 реализованных инструментов. Они делятся на несколько типов: преобразование текста в серию картинок (Story Picturing Engine), преобразование текста в серию ассоциированных картинок (Text-to-Picture Synthesis System), преобразование текста в набор статичных сцен (TEXT-TO-SCENE), преобразование текста в анимацию, в том числе с учетом положения объектов в пространстве (TEXT-TO-ANIMATION).

Точность генерации визуализации во многом зависит от степени развития NLP и нейросетей, а также вычислительных мощностей компьютера. Однако уже сейчас есть приемлемые результаты генерации, с которыми можно работать в индустрии кинематографа [19] и мультипликации (см., например, CRAFT [20], Storyboarder [21]).

В рамках настоящей работы были обозначены инструменты, которые доказывают жизнеспособность идеи генерации визуализации из текста, а также подсказывают пути развития инструментов ScriptViz [22], SceneMaker [23], CRAFT, Storyboarder.

ScriptViz генерирует трехмерную анимацию в режиме реального времени. Пользователь вводит текст, учитывая следующие ограничения: текст должен быть написан на английском, иметь понятные формулировки и хорошую структуру. Инструмент ограничен базой трехмерных моделей, в генерации участвуют готовые объекты и анимации.

SceneMaker – это инструмент, который автоматизирует генерацию ряда артефактов для кинопроизводства. Генерация сцен из сценария планировалась с точностью до мимики виртуальных актеров, освещения и аудиосопровождения. Проект не был реализован.

CRAFT – нейросеть, обученная на 25 184 фрагментах мультсериала Флинстоуны. Нейросеть способна сгенерировать новые серии этого сериала по текстовому описанию. В собранных сценах ещё встречаются ошибки, такие как неправильное наложение объектов или неправильный выбор анимаций, но в целом разработка интересна.

Storyboarder – инструмент для создания раскадровок от компании Wonder Unit. В данном инструменте есть функционал генерации статичных кадров на основе введенных текстовых запросов. Однако стоит отметить, что Storyboarder понимает ограниченный словарь слов и некоторые альтернативные названия.

4.2. Другие инструменты визуализации

Перечисленные выше инструменты подтверждают гипотезу о том, что техническую документацию игровых сценаристов возможно визуализировать программным путем. Однако ключевое отличие прототипирования повествования игр заключается в реализации возможности интерактивного взаимодействия, а также вариативности сюжета, чего не подразумевают эти инструменты, применимые скорее для кинематографа и мультипликации.

Были найдены инструменты, которые возможно адаптировать для разработки игр: StoryFlow [24], Machination Tool [25], Orange3 [26].

StoryFlow позволяет представить структуру в виде Yarn. Она визуализирует вариативность в происходящих событиях, что крайне полезно для красивой и наглядной демонстрации игровых событий, взаимодействия персонажей и определения временных промежутков.

Machination Tool – это приложение для моделирования и балансирования игровых систем. Данный инструмент позволяет визуализировать и симулировать игровые системы в виде динамических диаграмм. Данное решение может быть одним из элементов уже не сценарного, а скорее игрового прототипа.

Orange3 – это программный пакет визуального программирования на основе компонентов для визуализации данных, машинного обучения и интеллектуального анализа данных. В программу включены сотни готовых нод, каждая из которых отвечает за свою часть работы: различные визуализаторы (схемы, графики, таблицы), алгоритмы обработки и препроцессинга текста, готовые к обучению и работе нейросети. Помимо этого, разработчики предоставляют пользователям возможность реализовать собственный функционал в виде программ на Python, для которых есть отдельный нод.

5. ВИДЕНИЕ ИНСТРУМЕНТА ГЕНЕРАЦИИ СЦЕНАРНОГО ПРОТОТИПА

Для представления видения генератора сценарного прототипа разработана схема состава инструмента, а также проанализировано влияние наличия такого инструмента на процесс разработки видеоигр.

5.1. Схема и видение инструмента

На рис. 3 представлена схема инструмента, которая была разработана для последующей реализации.

Рис. 3. Схема инструмента генерации сценарного прототип

На вход инструмент принимает текст, который написан на естественном языке. Допускается, что документы будут удовлетворять некоторым правилам форматирования для лучшего распознания сущностей. В качестве текста подразумевается техническая документация сценариста, представленная в различных форматах.

Инструмент разделен на подсистемы: UI – пользовательский интерфейс и генератор сценарного прототипа.

Через пользовательский интерфейс пользователь будет иметь доступ к следующему функционалу, который разделен на модули:

- Модуль ввода принимает на вход текстовые документы и таблицы, а также другие форматы, в которых могут храниться данные: html, xml, json.
- Модуль редактирования позволяет вводить и редактировать текстовый документ, редактировать сгенерированный прототип и структуру повествования.
- Модуль отображения демонстрирует результаты визуализации структуры повествования, а также информацию об извлеченных сущностях (персонажах, локации и т. п.) и статистике по проекту (количество персонажей, локаций и т. п.).
- Модуль запуска позволяет пользователю запускать сценарный прототип, в который можно сыграть. Кроме этого возможен запуск сгенерированной балансной диаграммы.

Генератор сценарного прототипа разделен на модули по типам процессов:

- Модуль анализа текста анализирует входные данные с помощью алгоритмов NLP и машинного обучения. На выходе из него ожидаются перечень сущностей с типами, а также связи между сущностями.
- Модуль анализа данных и сборки структуры переводит информацию в понятный для алгоритмов генерации. На этом этапе обработки подсистема генерации сценарного прототипа включает в себя три модуля, каждый из которых последовательно обрабатывает результат работы предыдущих модулей:

a. Модуль анализа текста распознает необходимые сущности и связи между ними с помощью инструментов NLP;

b. Модуль анализа данных и построения структуры устанавливает целостную структуру сценария, связывает сущности между собой и строит структуру сценарного прототипа;

c. Модуль визуализации включает в себя четыре процесса: визуализация разветвленной структуры, постановка кадра, генерация окружающего пространства и генерация события в пространстве. Модуль получает на вход информацию в формализованном виде. На выходе выдает графические артефакты, ко-

торые можно воспроизвести в пользовательском интерфейсе. Кроме этого планируется, что в этом модуле будет происходить сборка балансных диаграмм на манер Machination.

Был составлен следующий план разработки инструмента генерации сценарного прототипа:

- визуализация структуры игрового сценария, представленного в виде текста на естественном языке (далее называемый TEXT) в виде графов или yarn-представлений;
- автоматическая сборка интерактивной раскадровки (для ограниченного количества жанров в виде визуальной новеллы или point-and-click игры) сценария игры из текста;
- создание алгоритма для генерации интерактивного сценарного прототипа из текста в игровом движке Unity;
- автоматическая генерация диаграммы баланса Machination на основе текста;
- автоматическая сборка игрового прототипа игры из текста.

Весь перечисленный функционал мы видим в рамках целостного приложения, на вход которому поступает серия текстовых документов, а на выходе получается проект, который объединяет в себе сгенерированный сценарный и, в перспективе, игровой прототип игры.

5.2. Влияние инструмента на разработку видеоигр

Процесс работы с повествовательной составляющей игры с применением нашего инструмента видится следующим:

- написание сценария игры в виде обычного текста на естественном языке в форме технической документации (текст, таблицы, wiki-разметка);
- генерация сценарного прототипа на основе документации;
- презентация и обсуждение сценарного прототипа внутри команды, с заказчиком и тестирование на фокус-группе;
- редактирование сценарного прототипа.

6. ПРОГРЕСС В РАЗРАБОТКЕ ИНСТРУМЕНТА

Решению задачи создания генератора сценарных и игровых прототипов посвящена серия наших разработок. В работе [27] были описаны собственный подход к созданию генератора сценарного прототипа и пилотная реализация инструмента для создания сцен на основе извлеченных из текста сущностей. В [28] представлена (см. рис. 4) программа для Orange 3, которая принимает на вход формализованный текст, а на выходе показывает различную визуализацию структуры: геометрическую связанность локаций, в каких локациях появляется персонаж, какие реплики произносятся в этих сценах и т. д.

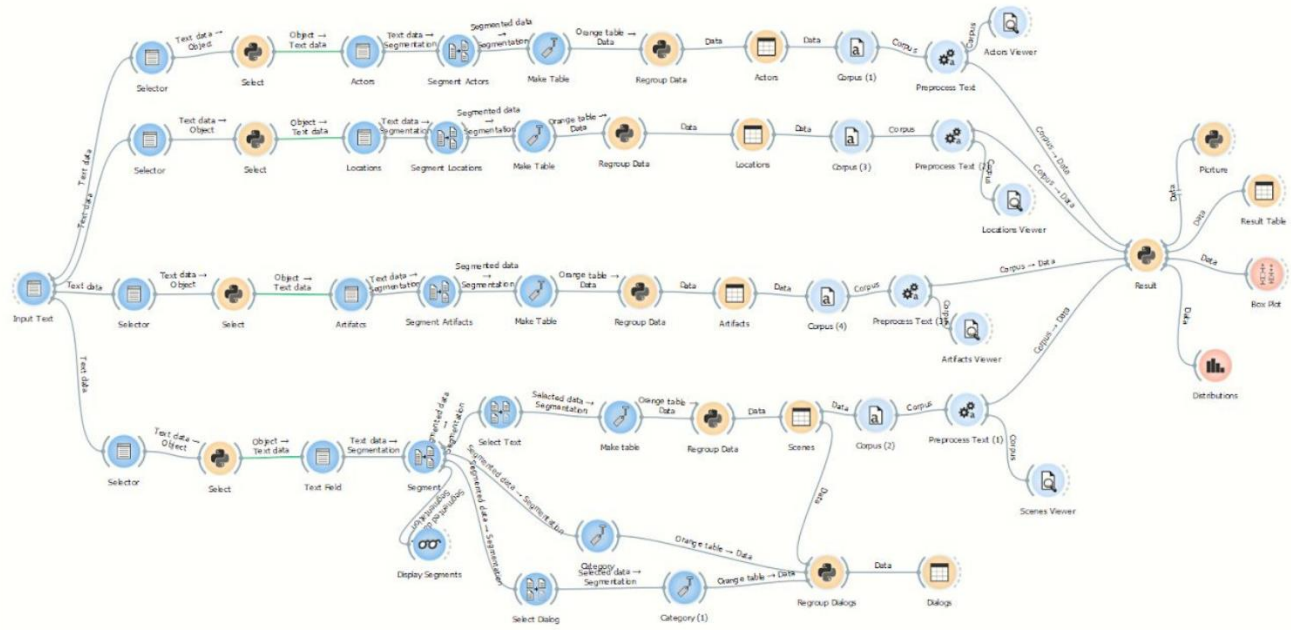


Рис. 4. Нодовая система генерации визуализации сценария

Другая работа (см., например, [29], [30]) связана с решением задачи извлечения информации о постановке кадра из текста. Для этого создана система анализа текста и вычисления необходимых настроек камеры из контекста. Работа находится на раннем этапе: воссоздается функционал текстового распознания кадра, аналогичный Storyboarder (см. рис. 5).

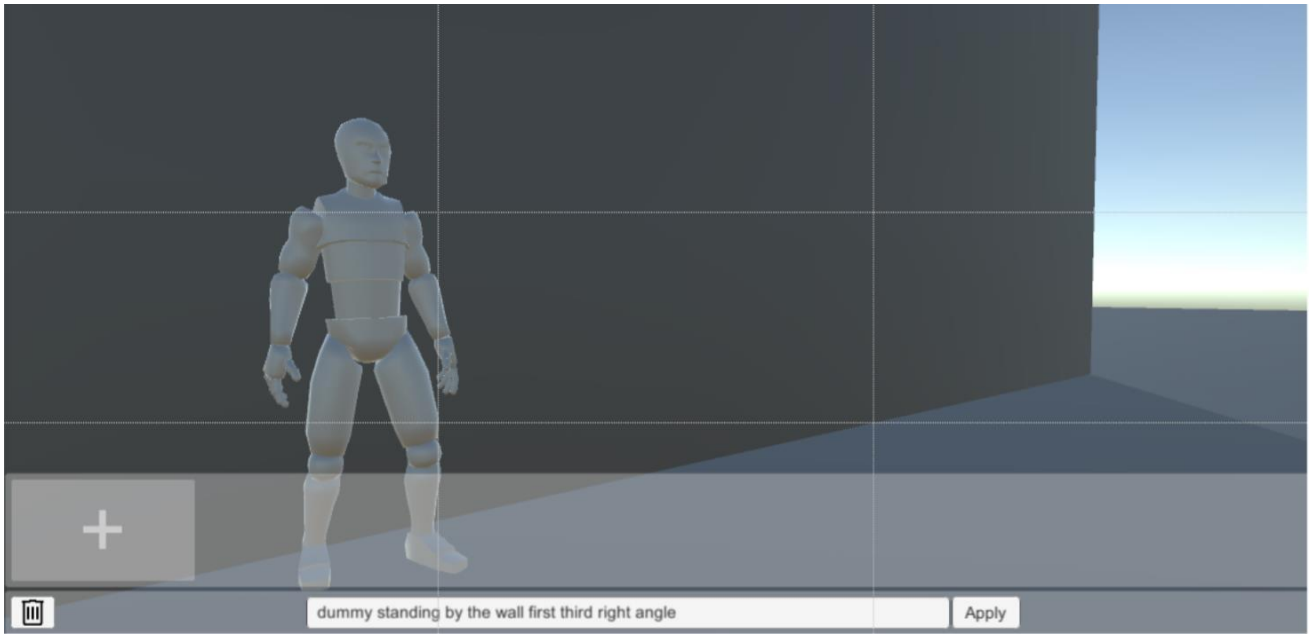


Рис. 5. Извлечение информации о кадре из текста

7. ВИДЕНИЕ ПРОЦЕССА ГЕНЕРАЦИЯ СОБЫТИЙ В ПРОСТРАНСТВЕ

Одна из задач модуля визуализации генератора сценарного прототипа – генерация визуализации событий в пространстве. Для этого необходима информация о персонажах в сцене, их действиях и репликах.

В рамках визуализации событий в пространстве персонажи должны расположиться в позах, соответствующим контексту из текста. На выходе должна получиться серия кадров, которая составляет из себя раскадровку. Данный шаг необходим как один из этапов разработки интерактивной раскадровки, в которой переход к следующему кадру будет зависеть от выбора игрока. В качестве платформы для экспериментов был выбран инструмент Storyboarder, который содержит необходимый функционал генерации визуализации.

Принцип работы состоит в следующем: на вход отправляется текст на английском языке, предложения которого представляют собой описание отдельных сцен. В каждом предложении есть информация о том, кто присутствует в сцене и что он делает. Текст анализируется и из него извлекается информация для визуализации – персонаж и его действие. Результат обрабатывается и отправляется на вход Storyboarder, где генерируются кадры, которые собираются в раскадровку для оценки результата.

Далее представлены подробности технической реализации и результаты эксперимента.

8. РЕАЛИЗАЦИЯ ПРОЦЕССА ГЕНЕРАЦИИ СОБЫТИЙ В ПРОСТРАНСТВЕ

Процесс генерации события в пространстве по текстовому вводу представляет собой автоматический конвейер. В нем происходят прием текста и его обработка. Затем результат отправляется на генерацию раскадровки в Storyboarder.

Возможность генерации элементов раскадровки по заданному вектору слов присутствует только в слепке коммита abbf5f24 открытого официального репозитория Storyboarder [31]. По этой причине в рамках эксперимента использована именно эта версия, а не последняя актуальная.

Для того чтобы интегрировать Storyboarder в конвейер, был изолирован функционал, который позволяет вызывать процедуру генерации с аргументом в виде текста и инициировать все необходимые механизмы для получения раскадровки. Вход – это текст, находящийся в управлении другого программного компонента, а выход – это сгенерированная раскадровка.

Также был внедрен микро веб-сервер express, который принимает на loopback интерфейсе запрос с параметром в виде входного текста, передаёт данный параметр в изолированный функционал и в качестве ответа отправляет тот текст, что получил, чтобы отправитель удостоверился в успешности операции генерации.

Исходный текст преобразуется в корпус [32] – это отобранная и обработанная по определенным правилам совокупность текстов, используемых в качестве базы для исследования языка. Он используются для статистического анализа и проверки статистических гипотез и подтверждения лингвистических правил в данном языке.

Так как машинное обучение в большинстве своем имеет дело с функционалами (любыми функциями, где образ является числом или множеством чисел), для обработки текста его приходится векторизовать. Этот процесс означает преобразование чисел в векторы по определенным правилам, таким, чтобы потеря информации, заложенной в тексте, была минимальной.

Также одним из этапов обработки текста является выделение стоп-слов – таких слов, которые в рамках генеральной совокупности играют малую роль в исследовании свойств языка. Например, в английском языке это будут слова “a”, “am”, “an”, “is”, “are” и т. д. В русском языке таковыми будут являться “же”, “то”, “бы” и т. д.

Для того чтобы Storyboarder корректно сгенерировал раскадровку, входной текст необходимо обработать. Первым этапом будет являться токенизация, а именно, преобразование текста в корпус в виде массива предложений; также каждое предложение необходимо разбить по словам и убрать пунктуацию. Во второй этап входит фильтрование всех слов в корпусе, в результате которого корпус будет очищен от стоп-слов. Далее стоит задача распознавания имён собственных и преобразования их в обезличенную женскую или мужскую сущность. Это необходимо сделать, так как Storyboarder не может различать имен, но может различать пол сущностей, преобразуя его в кадре в модель мужчины или женщины.

После проведения токенизации, во время этапа фильтрации, были использованы стоп-слова, собранные в приложении к работе [33]. Процесс фильтрации происходил итеративно, слова сравнивались посимвольно полно. На следующем этапе алгоритма необходимо построить классификатор, который будет распознавать в тексте имена собственные и определять их пол.

В процессе разработки в качестве модели выбран мультиклассовый наивный байесовский классификатор, так как закономерность выбора имен в обществе отсутствует и имена даются независимо друг от друга. Имена для обучения классификатора взяты из публичных данных Службы социального обеспечения США [34]. Для каждого из имени в обучающей выборке происходят векторизация в символьные биграммы, а также частотный анализ, исходя из которого виден процент использования имени в качестве женского и мужского. Так как обучение наивного байесовского классификатора – это лишь вычисление независимых вероятностей, произведение которых лежит в знаменателе формулы по теореме Байеса [35], трата времени на процесс минимальна.

Обученный байесовский классификатор отвечает нулём, если имя, пришедшее на вход, женское, и единицей, если мужское. Используя операцию классификации, каждый раз, когда в тексте появляется имя собственное, оно заменяется на слово man или woman в зависимости от пола, который присвоил ему обученный классификатор. Обезличенные слова выбраны как наиболее понятные для парсера предложений Storyboarder.

9. РЕЗУЛЬТАТ РАБОТЫ ПРОЦЕССА ВИЗУАЛИЗАЦИИ

9.1. Пример визуализации

Стоит отметить, что текст на естественном языке написан с оговоркой, что Storyboarder понимает ограниченный список слов. Важно, чтобы в сцене появлялся персонаж с анимацией, возможной в рамках Storyboarder. Настройки камеры задаются вручную последовательностью параметров, которые понимает Storyboarder. Итак, на вход отправляется следующий текст:

Bob is walking.

Alice says hi.

Bob is walking and looking back.

Bob is walking.

Alice hangs one arm.

Alice crossing hands.

Далее, в зависимости от контекста предложения, формулируются необходимые настройки камеры. Они хранятся в отдельном файле:

looking forward, medium long, single person, centered, left angle, eye level, long lens, light, frontrightlit, outside;

looking forward, medium long, single person, centered, right angle, eye level, long lens, light, frontrightlit, outside.

Список процессов преобразования, которые осуществляются над текстом:

1. Токенизация с очищением от пунктуации и стоп-слов.
2. Распознание и обезличивание имен собственных путем замены на слова man/woman в зависимости от пола имени с помощью наивного байесовского классификатора.

3. Полученный после преобразования текст дополняется настройками камер с помощью конкатенации строк.

Таким образом формируется результат, подающийся на вход в Storyboarder:

'man walking, looking forward, medium long, single person, centered, left angle, eye level, long lens, light, frontrightlit, outside';

'woman say hi, looking forward, medium long, single person, centered, right angle, eye level, long lens, light, frontrightlit, outside';

'man walk look back, looking forward, medium long, single person, centered, left angle, eye level, long lens, light, frontrightlit, outside';

'man walking, looking forward, medium long, single person, centered, left angle, eye level, long lens, light, frontrightlit, outside';

'woman hang one arm, looking forward, medium long, single person, centered, left angle, eye level, long lens, light, frontrightlit, outside';

'woman crossing hands, looking forward, medium long, single person, centered, left angle, eye level, long lens, light, frontrightlit, outside'.

Storyboarder генерирует раскадровку на основании полученных данных (см. рис. 6).

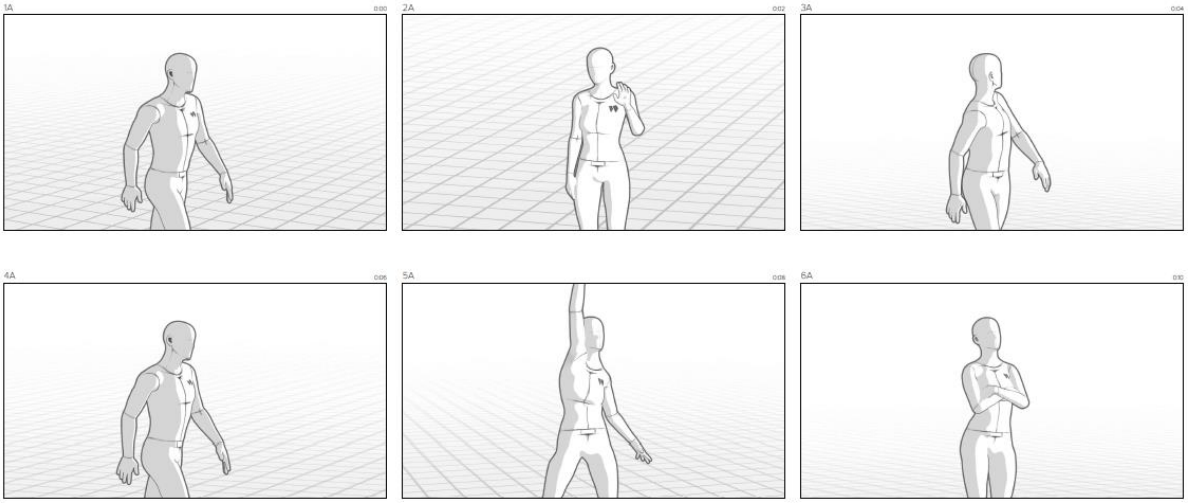


Рис. 6. Пример сгенерированной раскадровки

Возможности для генерации комикса в Storyboarder позволяют использовать его в случаях с подготовленным текстом. Расширение функционала и корпуса

Storyboarder позволит эффективнее работать с визуализацией натурального текста.

В дальнейшем планируется разработка собственного инструмента, а не доработка существующих решений.

9.2. Эксперимент

В качестве эксперимента был взят сценарий существующего проекта [36]. Сценарий был переписан в вид, в котором есть вся необходимая информация для генерации кадра в Storyboarder. Данный текст был предоставлен для ручной сборки раскадровки в Storyboarder. Настройки камеры использовались такие же, как и для генерации раскадровки.

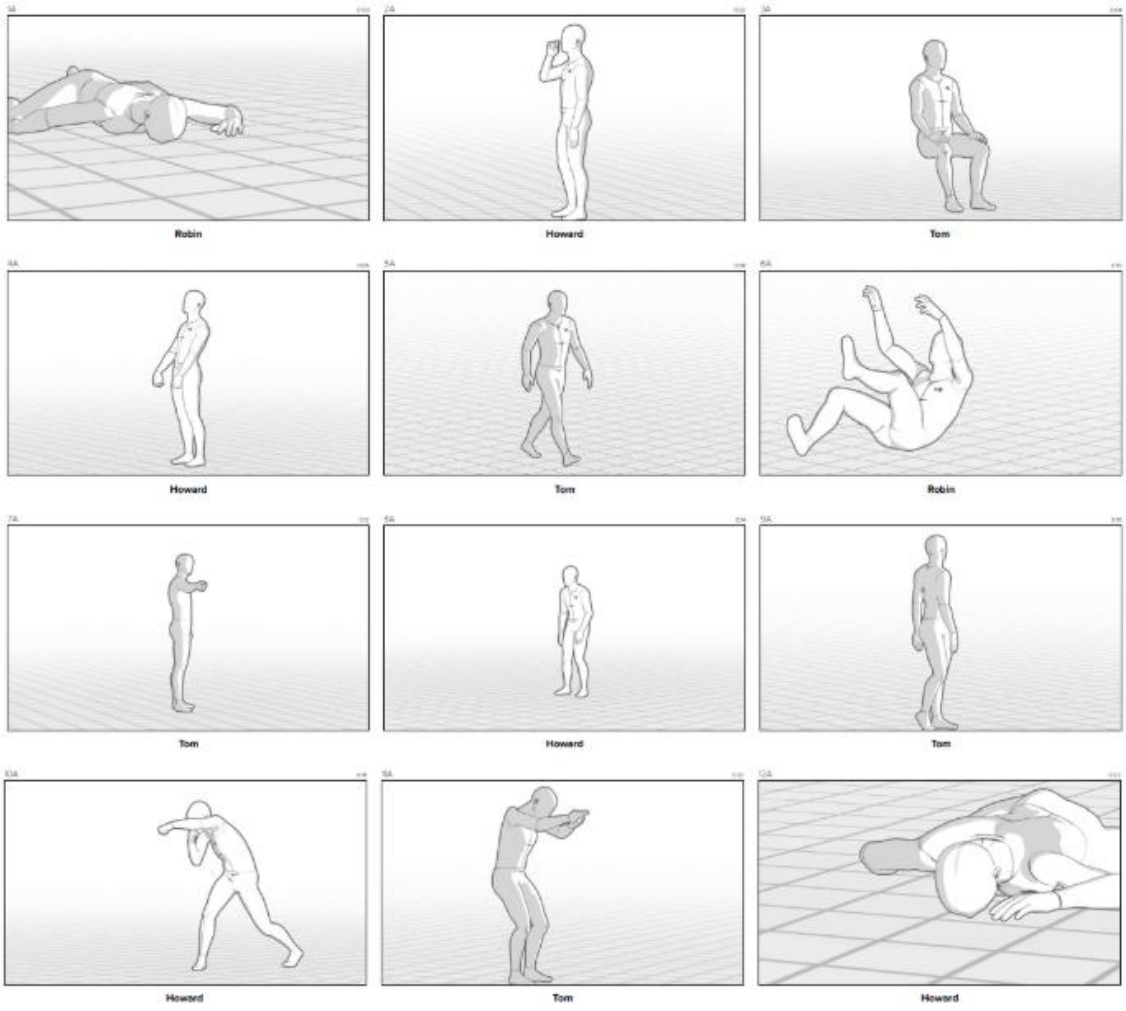


Рис. 7. Результат генерации для эксперимента

Вместе с воспроизведением процесс ручного создания раскадровки занял 236 секунды или 3 минуты и 56 секунд. Далее для сравнения текст был отправлен на автоматическую генерацию кадров. Раскадровка идентична той, что получена вручную (см. рис. 7). Вместе с воспроизведением процесс занят 42 секунды, что в 5,6 раз быстрее ручной сборки. Стоит отметить, что чем больше проект, тем дольше собирать его вручную.

Гипотеза о том, что инструмент для генерации раскадровок является одним из эффективных инструментов для создания визуализации сценарного прототипа, подтверждается.

10. ДАЛЬНЕЙШЕЕ РАЗВИТИЕ ИНСТРУМЕНТА

Сложность дальнейшего развития инструмента заключается в том, что интерактивное повествование по умолчанию разветвленное и содержит в себе возможность выбора. Задача извлечения подобных зависимостей реальна при применении технологий машинного обучения. Для этого обучения нейросетей необходимы корпуса, которых на данный момент не существует.

В дальнейшем разработка должна перейти с прототипных и экспериментальных решений в форму самостоятельного инструмента, разработанного с нуля.

В ходе обсуждения результатов разработки были сформулированы задачи на будущее:

- Извлечение информации о постановке кадра из текста;
- Разметка и форматирование технической документации;
- Создание интерактивной раскадровки;
- Генерация визуализации.

Нереализованной осталась задача постановки кадра в зависимости от контекста повествования. Данный вопрос важен с точки зрения автоматизации генерации визуализации сценария, и уже начата разработка в этом направлении.

Мы не касались темы классической разметки сценарных документов для неинтерактивных произведений и особенностей сценариев для интерактивных произведений. Данному вопросу мы планируем посвятить время в дальнейшем. Решение вопроса разметки и форматирования документов с нелинейным повествованием поможет решить следующую задачу, связанную с интерактивностью раскадровки.

Раскадровки хороши в кинематографе и анимации, а также в играх, например, в создании кат-сцен. Главное изменение, которое мы хотим внести в покадровое представление игрового повествования, – это возможность переключаться с помощью взаимодействия с игровой действительностью. Интерактивная раскадровка отдаленно напоминает нам визуальную новеллу или игры жанра point-and-click. В игре повествование движется за счет действий игрока. Исключая из прототипа влияние геймплея и выбор действий, мы лишаем сценаристов возможности тестирования повествования в условиях реальной разработки. В сценарном прототипе интерактив может быть заменен геймплейными “заглушками”, которые однозначно трактуют одну из веток развития событий. С таким инструментом игровой сценарист и нарративный дизайнер смогут заранее «отыграть» сюжет игры, показать его разработчикам, заказчику и пользователям. Возможность оценки и тестирования сценарного прототипа поможет оценить черты игр, как классические для линейного повествования, так и характерные для интерактивного повествования: погружение, темпоритм, формирование эмпатии. Таким образом, следующей задачей мы ставим приведение раскадровки к интерактивному виду.

Вопрос о генерации визуализации по текстовому описанию связан с тем, что необходимо заранее создать необходимые трехмерные объекты, анимации для них, логику сборки сцены, непротиворечивые правила взаимодействия между объектами. При этом каждая из перечисленных единиц сборки должна иметь свойство, схожее со свойством вектора в базисе линейного пространства. Другими словами, количество созданных единиц должно быть минимальным при условии, что вариативность генерируемых сцен должна быть максимальной. Данный момент необходимо учитывать в последующих разработках.

После усовершенствования прототипа планируется продолжить тестирование, а также сформулировать метрики, которые покажут, насколько генерация сценарного прототипа эффективнее ручного прототипирования.

ЗАКЛЮЧЕНИЕ

В работе введен термин сценарного прототипа и описано авторское представление о нём. Помимо этого, сформулирована актуальность инструмента генерации сценарного прототипа.

В работе приведен анализ инструментов, которые вдохновили авторов. Существование этих инструментов говорит о том, что разработка инструмента генерации сценарного прототипа возможна. Данный факт подтверждают примеры инструментов коллег, которые были разработаны с участием авторов статьи (см. в разделе 6).

В качестве эксперимента был разработан прототип компонента визуализации инструмента генерации сценарного прототипа. Прототип генерирует раскадровку на основе текста. Компонент призван сэкономить время на визуализацию игрового повествования.

В результате проделанной работы были детализированы требования к инструменту генерации. Размышления на тему развития инструмента привели к формированию ряда задач для следующей итерации прототипирования.

Данная разработка интересна в рамках использования для разработки не только игр, но и serious games, симуляторов и виртуальных обучающих тренажеров, которым также посвящено наше пристальное внимание (см., например, [36–38]).

СПИСОК ЛИТЕРАТУРЫ

1. Open-source tool for telling interactive, nonlinear stories Twine. URL: https://twinery.org/.

2. Visual novel engine Ren'Py. URL: https://www.renpy.org/.

3. The solution for interactive storytelling and game content management articy:draft 3. URL: https://www.articy.com/en/.

4. Real-time 3D development platform Unity. URL: https://unity.com/.

5. Real-time 3D creation tool Unreal Engine. URL: https://www.unrealengine.com/en-US/.

6. An Online Visual Collaboration Platform for Teamwork Miro. URL: https://miro.com/.

7. Diagram Software and Flowchart Make diagrams.net.
URL: https://www.diagrams.net/.

8. *Davies D., Bathurst D., Bathurst R.* The Telling Image: The Changing Balance between Pictures and Words in a Technological Age // Technology and Culture. 1992. No. 4. P. 845–846.

9. *Raja D., Bowman D., Lucas J., North C.* Exploring the benefits of immersion in abstract information visualization // Proc. of the 8th Int'l Immersive Projection Technology Workshop. 2004.
URL: https://people.cs.vt.edu/bowman/papers/ipt_dheva.pdf.

10. *Montfort N., Bogost I.* Racing the Beam: The Atari Video Computer System. Cambridge: MIT Press, 2009. 192 p.

11. *Bogost I.* Videogames are a Mess.
URL: http://bogost.com/writing/videogames_are_a_mess/.

12. *Cairns P., Cox A., Nordin I.* Immersion in Digital Games: Review of Gaming Experience Research // Handbook of Digital Games. 2014. P. 337–361.

13. *Newman J.* Videogames. Routledge: Psychology Press, 2014. 198 p.

14. *lanchard L.* Creating empathy in video games. The University of Dublin, 2016. 42 p.

15. *Koivisto E., Suomela R.* Using prototypes in early pervasive game development // Computers in Entertainment. 2008. No. 17.

16. Open-source plugin for Unity Cradle.
URL:https://forum.unity.com/threads/released-cradle-play-twine-stories-in-unity.333720/.

17. Point-and-click. URL: https://dic.academic.ru/dic.nsf/ruwiki/1443730.

18. *Hassani K., Lee W.-S.* Visualizing Natural Language Descriptions: A Survey // ACM Computing Surveys, 2016. URL: https://arxiv.org/pdf/1607.00623.pdf.

19. Artificial intelligence tools RivetAI. URL: https://www.rivetai.com/.

20. *Gupta T., Schwenk D., Farhadi A., Hoiem D., Kembhavi A.* Imagine This! Scripts to Compositions to Videos. Cornell University, 2018.
URL: https://openaccess.thecvf.com/content_ECCV_2018/papers/Tanmay_Gupta_Imagine_This_Scripts_ECCV_2018_paper.pdf.

21. Storyboarder. URL: https://wonderunit.com/storyboarder/.

22. *Liu Z.-Q., Leung K.-M.* Script visualization (ScriptViz): A smart system that makes writing fun. Switzerland: Springer, 2006. P. 34–40.

23. *Akser M., Bridges B., Campo G., Cheddad A., Curran K., Fitzpatrick L., Hamilton L., Harding J., Leath T., Lunney T., Lyons F., Ma M., Macrae J., Maguire T., McCaughey A., McClory E., McCollum V., Mc Kevitt P., Melvin A., Moore P., Mulholland E., Muñoz K., O'Hanlon G., Roman L.* SceneMaker: Creative technology for digital storytelling. Switzerland: Springer, 2016. P. 29–38.

24. *Padia K., Bandara K., Healey C.* A system for generating storyline visualizations using hierarchical task network planning // Computers & Graphics. 2019. V. 78. P. 64–75.

25. *Adams E., Joris D.* The Designer's Notebook: Machinations, A New Way to Design Game Mechanics.
URL: https://www.gamasutra.com/view/feature/176033/the_designers_notebook.

26. Orange 3. URL: https://orange.biolab.si/.

27. *Сахибгареева Г.Ф., Кугуракова В.В.* Концепт инструмента автоматического создания сценарного прототипа компьютерной игры // Электронные библиотеки. 2018. Т. 21. № 3-4. С. 235–249.

28. *Доброквашина А.С., Газизова Э.А.* Автоматизация проектирования игрового прототипа на основании обработки формализованного игрового дизайн-документа // Ученые записки ИСГЗ. 2019. Т. 17. № 1. С. 583–589.

29. *Астафьев А.М.* Разработка инструмента для сборки сцен по тегам // Казанский (Приволжский) федеральный университет, 2019. 31 с.
URL: https://kpfu.ru/student_diplom/10.160.178.20_FPEBER9KDIZQVYJAE3VRTIFYWZB_CDDM972OPP2I28S0EEFABT_Astafev.pdf.

30. *Кугуракова В.В., Сахибгареева Г.Ф., Нгуен А.З., Астафьев А.М.* Пространственная ориентация объектов на основе обработки текстов на естественном языке для генерации раскадровок // Электронные библиотеки. 2020. Т. 23. № 6. С. 1214–1238.

31. GitHub Storyboarder. URL: https://github.com/wonderunit/storyboarder.

32. *Николаев И.С., Митренина О.В., Ландо Т.М.* Прикладная и компьютерная лингвистика. М.: URSS, 2016. 320 с.

33. *Nothman J., Qin H., Yurchak R.* Stop Word Lists in Free Open-source Software Packages // Proc. Workshop for NLP Open Source Software. 2018. P. 712.

34. Baby Names from Social Security Card Applications. URL: https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-level-data.

35. *McCreery C.* First-year Statistics for Psychology Students Through Worked Examples. 1. Probability and Bayes' Theorem. Oxford Forum, 2018. 29 p.

36. *Антонов И.О., Зезегова К.В., Кугуракова В.В., Лазарев Е.Н., Хафизов М.Р.* Программирование запахов для виртуального осмотра места происшествия // Электронные библиотеки. 2018. Т. 21. № 3-4. С. 301–313.

37. *Abramov V., Kugurakova V., Rizvanov A., Abramskiy M., Manakhov N., Evstafiev M., Ivanov D.* Virtual Biotechnological Lab Development // BioNanoScience. 2017. V. 7. No. 2. P. 363–365.

38. *Kugurakova V., Abramov V., Sultanova R., Tsivilskiy I., Talanov M.* Virtual Reality-Based Immersive Simulation for Invasive Surgery Training // European Journal of Clinical Investigation. 2018. V. 48. P. 224–225.

STORYBOARD AS ONE OF THE REPRESENTATIONS OF THE SCENARIO PROTOTYPE OF COMPUTER GAMES

G. F. Sahibgareeva[1, [0000-0003-4673-3253]], O. A. Bedrin[2, [0000-0003-3300-4318]], V. V. Kugurakova[3, [0000-0002-1552-4910]]

[1, 2, 3]*Institute of ITIS, Kazan Federal University*

[1]gulnara.sahibgareeva42@gmail.com, [2]simplavero@gmail.com, [3]vlada.kugurakova@gmail.com

Abstract

The work is devoted to the study and improvement of the design, development, and testing of video game storytelling. The existing practices of writing and keeping up-to-date scripts for interactive works have been studied. The definition of a scenario

prototype and requirements for its form are formulated. An idea was put forward about the efficiency of automating the creation of a scenario prototype in the form of a generator tool. A vision of such a tool has been drawn up. The impact of such a tool on development order is presented. Implemented a tool component and conducted an experiment that proves its effectiveness with an example such as generating storyboards from the text. Plans for future development have been formulated.

Keywords*: computer games, video game development, interactive storytelling, scenario prototype, narrative design, screenwriting, game documentation, storyboarding, storyboard generation, interactive storyboarding.*

REFERENCES

1. Open-source tool for telling interactive, nonlinear stories Twine. URL: https://twinery.org/.

2. Visual novel engine Ren'Py. URL: https://www.renpy.org/.

3. The solution for interactive storytelling and game content management articy:draft 3. URL: https://www.articy.com/en/.

4. Real-time 3D development platform Unity. URL: https://unity.com/.

5. Real-time 3D creation tool Unreal Engine. URL: https://www.unrealengine.com/en-US/.

6. An Online Visual Collaboration Platform for Teamwork Miro. URL: https://miro.com/.

7. Diagram Software and Flowchart Make diagrams.net. URL: https://www.diagrams.net/.

8. *Davies D., Bathurst D., Bathurst R.* The Telling Image: The Changing Balance between Pictures and Words in a Technological Age // Technology and Culture. 1992. No. 4. P. 845–846.

9. *Raja D., Bowman D., Lucas J., North C.* Exploring the benefits of immersion in abstract information visualization // Proc. of the 8th Int'l Immersive Projection Technology Workshop. 2004. URL: https://people.cs.vt.edu/bowman/papers/ipt_dheva.pdf.

10. *Montfort N., Bogost I.* Racing the Beam: The Atari Video Computer System. Cambridge: MIT Press, 2009. 192 p.

11. *Bogost I.* Videogames are a Mess.
URL: http://bogost.com/writing/videogames_are_a_mess/.

12. *Cairns P., Cox A., Nordin I.* Immersion in Digital Games: Review of Gaming Experience Research // Handbook of Digital Games. 2014. P. 337–361.

13. *Newman J.* Videogames. Routledge: Psychology Press, 2014. 198 p.

14. *lanchard L.* Creating empathy in video games. The University of Dublin, 2016. 42 p.

15. *Koivisto E., Suomela R.* Using prototypes in early pervasive game development // Computers in Entertainment. 2008. No. 17.

16. Open-source plugin for Unity Cradle.
URL:https://forum.unity.com/threads/released-cradle-play-twine-stories-in-unity.333720/.

17. Point-and-click. URL: https://dic.academic.ru/dic.nsf/ruwiki/1443730.

18. *Hassani K., Lee W.-S.* Visualizing Natural Language Descriptions: A Survey // ACM Computing Surveys, 2016. URL: https://arxiv.org/pdf/1607.00623.pdf.

19. Artificial intelligence tools RivetAI. URL: https://www.rivetai.com/.

20. *Gupta T., Schwenk D., Farhadi A., Hoiem D., Kembhavi A.* Imagine This! Scripts to Compositions to Videos. Cornell University, 2018.
URL: https://openaccess.thecvf.com/content_ECCV_2018/papers/Tanmay_Gupta_Imagine_This_Scripts_ECCV_2018_paper.pdf.

21. Storyboarder. URL: https://wonderunit.com/storyboarder/.

22. *Liu Z.-Q., Leung K.-M.* Script visualization (ScriptViz): A smart system that makes writing fun. Switzerland: Springer, 2006. P. 34–40.

23. *Akser M., Bridges B., Campo G., Cheddad A., Curran K., Fitzpatrick L., Hamilton L., Harding J., Leath T., Lunney T., Lyons F., Ma M., Macrae J., Maguire T., McCaughey A., McClory E., McCollum V., Mc Kevitt P., Melvin A., Moore P., Mulholland E., Muñoz K., O'Hanlon G., Roman L.* SceneMaker: Creative technology for digital storytelling. Switzerland: Springer, 2016. P. 29–38.

24. *Padia K., Bandara K., Healey C.* A system for generating storyline visualizations using hierarchical task network planning // Computers & Graphics. 2019. V. 78. P. 64–75.

25. *Adams E., Joris D.* The Designer's Notebook: Machinations, A New Way to Design Game Mechanics.
URL: https://www.gamasutra.com/view/feature/176033/the_designers_notebook.

26. Orange 3. URL: https://orange.biolab.si/.

27. *Sahibgareeva G.F., Kugurakova V.V.* Koncept instrumenta avtomatich-eskogo sozdanija scenarnogo prototipa komp'juternoj igry // Jelektronnye biblioteki. 2018. T. 21. № 3-4. S. 235–249.

28. *Dobrokvashina A.S., Gazizova Je.A.* Avtomatizacija proektirovanija igrovogo prototipa na osnovanii obrabotki formalizovannogo igrovogo dizajn-doku-menta // Uchenye zapiski ISGZ. 2019. T. 17. № 1. S. 583–589.

29. *Astaf'ev A.M.* Razrabotka instrumenta dlja sborki scen po tegam. Kazanskij (Privolzhskij) federal'nyj universitet, 2019. 31 s.
URL: https://kpfu.ru/student_diplom/10.160.178.20_FPEBER9KDIZQVYJAE3VRTI-FYWZB_CDDM972OPP2I28S0EEFABT_Astafev.pdf.

30. *Kugurakova V.V., Sahibgareeva G.F., Nguen A.Z., Astaf'ev A.M.* Pros-transtvennaja orientacija ob#ektov na osnove obrabotki tekstov na estestvennom jazyke dlja generacii raskadrovok // Jelektronnye biblioteki. 2020. T. 23. № 6. S. 1214–1238.

31. GitHub Storyboarder. URL: https://github.com/wonderunit/storyboarder.

32. *Nikolaev I.S., Mitrenina O.V., Lando T.M.* Prikladnaja i komp'juternaja lingvistika. M.: URSS, 2016. 320 s.

33. *Nothman J., Qin H., Yurchak R.* Stop Word Lists in Free Open-source Soft-ware Packages // Proc. Workshop for NLP Open Source Software. 2018. P. 712.

34. Baby Names from Social Security Card Applications.
URL: https://catalog.data.gov/dataset/baby-names-from-social-security-card-applica-tions-national-level-data.

35. *McCreery C.* First-year Statistics for Psychology Students Through Worked Examples. 1. Probability and Bayes' Theorem. Oxford Forum, 2018. 29 p.

36. *Antonov I.O., Zezegova K.V., Kugurakova V.V., Lazarev E.N., Hafizov M.R.* Programmirovanie zapahov dlja virtual'nogo osmotra mesta proisshestvija // Jelektronnye biblioteki. 2018. T. 21. № 3-4. S. 301–313.

37. *Abramov V., Kugurakova V., Rizvanov A., Abramskiy M., Manakhov N., Evstafiev M., Ivanov D.* Virtual Biotechnological Lab Development // BioNanoScience. 2017. V. 7, No. 2. P. 363–365.

38. *Kugurakova V., Abramov V., Sultanova R., Tsivilskiy I., Talanov M.* Virtual Reality-Based Immersive Simulation for Invasive Surgery Training // European Journal of Clinical Investigation. 2018. V. 48. P. 224–225.

СВЕДЕНИЯ ОБ АВТОРАХ

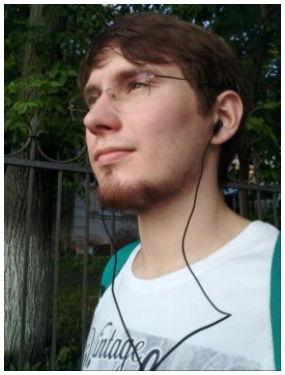
КУГУРАКОВА Влада Владимировна – к. т. н., доцент кафедры программной инженерии Института ИТИС КФУ, руководитель НИЛ разработки AR/VR приложений и компьютерных игр. Сфера научных интересов – иммерсивность виртуальных сред, проблемы генерации реалистичной визуализации, различные аспекты проектирования игр, AR/VR, подходы к интерпретации UX.

Vlada Vladimirovna KUGURAKOVA, PhD., Docent of the Institute ITIS KFU, Head of Laboratory «AR/VR applications and Gamedev». Research interests include immersiveness of virtual environments, problems of generating realistic visualization, various aspects of game design, AR/VR, approaches to UX interpretation.

vlada.kugurakova@gmail.com

САХИБГАРЕЕВА Гульнара Фаритовна – ассистент кафедры программной инженерии Института ИТИС КФУ. Сфера научных интересов – игровая сценаристика, нарративный дизайн, изучение вопроса эффективности создания сценарного прототипа и возможности автоматизации данного процесса.

Gulnara Faritovna SAHIBGAREEVA – assistant of the Department of Software Engineering of the Institute ITIS KFU. Research interests - game scripting, narrative design, studying the issue of the effectiveness of creating a scenario prototype and the possibility of automating this process. gulnara.sahibgareeva42@gmail.com



БЕДРИН Олег Александрович – бакалавр Института ИТИС КФУ. Сфера научных интересов – машинное обучение, разработка видеоигр.

Oleg Aleksandrovich BEDRIN – bachelor of the Institute ITIS KFU. Research interests – machine learning, development.

simplavero@gmail.com

Материал поступил в редакцию 17 ноября 2020 года