УДК 004.4

ИЗВЛЕЧЕНИЕ ЗНАНИЙ ИЗ WIKIDATA ДЛЯ ФОРМИРОВАНИЯ МЕТАДАННЫХ ДОКУМЕНТОВ ЭЛЕКТРОННЫХ МАТЕМАТИЧЕСКИХ КОЛЛЕКЦИЙ

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

[1–3] *Институт информационных технологий и интеллектуальных систем Казанского федерального университета, ул. Кремлевская, 35, г. Казань, 420008*

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

Аннотация

Представлены методы создания цифровых математических коллекций, включающих неструктурированные наборы документов. Эти наборы содержат материалы сборников научных конференций, а также статьи из архивов математических журналов «доцифрового» периода.

Формирование обязательного набора метаданных названных документов произведено с помощью программных инструментов фабрики метаданных цифровой математической библиотеки Lobachevskii DML. Для уточнения и пополнения наборов метаданных документов цифровых коллекций использованы методы извлечения знаний из Wikidata.

Разработана система SPARQL-запросов для поиска в Wikidata информации о документах электронных коллекций и их авторах. Обозначен набор сущностей Wikidata, определяющих признаки поиска, а также последующую фильтрацию полученных результатов.

Предложены методы уточнения и дополнения библиографических ссылок, приведенных в статьях. При формировании метаданных документов ретро-коллекций произведен поиск в Wikidata сведений о годах жизни авторов статей, а также URL веб-страниц с информацией о статьях и их авторах. Приведены результаты формирования нескольких новых электронных коллекций цифровой библиотеки Lobachevskii-DML.

***Ключевые слова**: Wikidata, метаданные, фабрика метаданных, цифровая математическая коллекция, цифровая математическая ретро коллекция, цифровая математическая библиотека, Lobachevskii-DML.*

ВВЕДЕНИЕ

В настоящее время происходящие изменения в инфраструктуре научных коммуникаций поставили целый ряд новых задач по управлению знаниями, а каждый этап жизненного цикла научной публикации предполагает его сопровождение специализированными программными инструментами (например, [1–4]).

Сегодня необходимым элементом научного исследования стало описание связей новых научных результатов с полученными ранее, что может быть выполнено в современных условиях только при наличии в Сети научного контента за весь период исследования рассматриваемых научных проблем. Такие связи устанавливаются все более активно во всех научных дисциплинах, поэтому можно утверждать, что в настоящее время формируется общее пространство научных знаний (см., например, [5]). В частности, основные направления интеграции математического знания определены в проектах “The Global Digital Mathematics Library” (GDML) и “World Digital Mathematics Library” (WDML) [6–8].

Метаданные являются основой коммуникации в Сети и используются на всех этапах жизненного цикла научной публикации (например, [9]). В настоящее время все научные документы оформляются для публикации с помощью программных инструментов – в англоязычной научной литературе для обозначения этого процесса используется термин “born-digital” (см. [10]). Современные правила подготовки научных публикаций, как в специализированных научных журналах, так и в сетевых изданиях, содержат требования по включению в соответствующие документы предметных классификаторов, ключевых слов, ORCID авторов и некоторой другой информации (например, [11, 12]). Именно на основании этой информации формируется набор метаданных научного документа.

При создании электронных коллекций научных документов, изданных в «доцифровой» период, возникают определенные проблемы с формированием обязательного набора метаданных документа (см., например, [13]). В такой ситуации с помощью методов анализа структуры и стилевых особенностей документа

можно сформировать основной набор его метаданных, включающий название этого документа, список его авторов и библиографию [14–17].

Ключевые слова и предметные классификаторы, такие как УДК [18] и Mathematics Subject Classification (MSC) [19], являются обязательными атрибутами современной научной публикации. Для создания или расширения списка ключевых слов используются методы текстового анализа. Подбор предметных классификаторов для математических статей производится методами автоматической классификации и категоризации (например, [20–23]). Но этих методов недостаточно для получения полного набора метаданных. Например, при формировании научных ретро-коллекций возникают проблемы даже с получением полной информации об авторах документов. Отметим, что новые методы формирования метаданных математических документов разрабатываются в проектах создания цифровых математических библиотек (см., например, [24, 25]).

На протяжении последних лет новые электронные математические коллекции формируются нами в рамках проекта создания цифровой библиотеки Lobachevskii Digital Mathematical Library (Lobachevskii-DML, https://lobachevskii-dml.ru/). Основной целью этого проекта является построение системы взаимосвязанных программных сервисов, обеспечивающих создание, обработку, хранение и управление объектами цифровых библиотек, а также интеграцию создаваемых электронных коллекций в агрегирующие цифровые математические библиотеки [26–28].

В настоящей работе представлен метод создания обязательного набора метаданных документов ретро-коллекций цифровой математической библиотеки. Термин «обязательный набор метаданных» понимается нами в соответствии со схемой метаданных EuDML [29]. В качестве источника пополнения метаданных использованы открытые ресурсы Веба. С помощью программных инструментов фабрики метаданных цифровой математической библиотеки Lobachevskii-DML реализованы основные процессы текстового анализа документов электронных ретро-коллекций, в частности, выделение именованных сущностей. С помощью разработанной системы запросов произведен поиск в Сети информации, необходимой для получения метаданных, с последующей экстракцией информационных объектов. После автоматизированного проведения фильтрации и нормали-

зации полученная информация включается в набор метаданных. Как один из основных результатов, представлен процесс формирования обязательного набора метаданных ретро-коллекции статей журнала «Известия физико-математического общества при Казанском университете» – одной из электронных коллекций цифровой библиотеки Lobachevskii-DML.

В разделе 1 выделены основные процессы создания математических электронных коллекций для их включения в цифровые библиотеки. Отмечены особенности формирования метаданных документов таких коллекций в соответствии со схемами агрегирующих цифровых библиотек.

В разделе 2 выделены наиболее важные проблемы формирования метаданных документов электронных математических ретро-коллекций.

Третий раздел посвящен методам получения информации из Wikidata с целью пополнения и уточнения метаданных документов электронных коллекций.

В четвертом разделе приведены алгоритмы пополнения метаданных документов ретро-коллекций цифровой библиотеки Lobachevskii-DML информацией, полученной из Wikidata.

1. ЦИФРОВЫЕ МАТЕМАТИЧЕСКИЕ БИБЛИОТЕКИ КАК ЧАСТЬ СПЕЦИАЛИЗИРОВАННОЙ НАУЧНОЙ ИНФРАСТРУКТУРЫ

Как отмечено в [8, 24], цифровым математическим библиотекам отводится роль основного интегратора математического знания, представленного в научных документах, опубликованных когда-либо. Обзор наиболее значительных цифровых математических библиотек приведен в [25, 30].

Проблемы интеграции знаний, полученных в области математики за весь «печатный» период развития этой науки, рассматривались в целом ряде проектов. Даже если эти проекты носили локальный характер, методы и инструменты, разрабатываемые в ходе их выполнения, были ориентированы на всеобъемлющую интеграцию математических знаний (см., например, [24]), а достигнутый уровень развития позволил поставить вопрос создания Всемирной цифровой математической библиотеки WDML.

Цифровая библиотека Lobachevskii DML включает в себя ряд электронных коллекций, при создании которых было необходимо выполнить полный цикл их

формирования: от оцифровки бумажных документов до загрузки цифровых документов и их метаданных в библиотеку. К числу таких коллекций относятся, например, «Труды Математического центра им. Н.И. Лобачевского» (далее – «Труды …» [31], а также ретро-коллекция «Известия физико-математического общества при Казанском университете» (“Bulletin de la Société Physico-Mathémaique de Kasan”) (далее – «Известия …» [32]. «Труды …» издаются с 1998 года, а до 2015 года большая часть их выпусков была только на бумажных носителях. Архивы журнала «Известия …» хранятся в Научной библиотеке Казанского университета только в бумажном виде и, как правило, в единичных экземплярах.

Создание электронной коллекции математических документов состоит из следующих основных этапов:

- Сканирование и оптическое распознавание документов;
- Разбиение архива документов на группы на основании сходства структуры и стилевого оформления документов;
- Определение начальной и завершающей страниц статей в файлах каждой группы документов;
- Разделение файлов на отдельные статьи на основании данных, полученных на предыдущем этапе;
- Поиск и выделение из документов основных метаданных;
- Уточнение метаданных;
- Поиск и пополнение метаданных информацией из Wikidata;
- Формирование метаданных статей по xml-схемам цифровой библиотеки;
- Интеграция электронной коллекции в соответствующую цифровую математическую библиотеку;
- Нормализация метаданных по xml-схемам агрегирующих цифровых библиотек.

Метаданные документов электронных коллекций, представленных в настоящей статье, были сформированы программными сервисами фабрики метаданных цифровой библиотеки Lobachevskii-DML [28, 33]. Эти сервисы реализуют методы, основанные на анализе структуры документов и особенностях их стилевого оформления [14, 16]. В основе реализации этих инструментов лежат методы ана-

лиза структуры документов (см., например, [14–17, 34, 35]). Также при формировании метаданных были применены стандартные алгоритмы текстового анализа (см., например, [36–38]).

Особенностью электронной коллекции «Труды …», как и многих других сборников материалов конференций, является отсутствие единых изначально сформулированных требований к структуре научных документов, включенных в эти издания. Это обстоятельство усложняет процесс извлечения метаданных методами, основанными на анализе структуры документа и его стилевых признаков. Так, например, ключевые слова, аннотации и предметные классификаторы присутствуют в статьях лишь в незначительном количестве сборников, в то время как эта информация необходима для формирования наборов метаданных по схемам агрегаторов математических документов [29, 39].

Далее, с использованием инструментов фабрики метаданных цифровой библиотеки Lobachevskii-DML нами была проведена процедура нормализации метаданных в соответствии с DTD-правилами и XML-схемами Journal Archiving and Interchange Tag Suite (NISO JATS) [40]. Для этого был сформирован набор метаданных в виде item-структуры, включающей как содержание метаданных, так и информацию о языке их представления. Это позволило включить в набор метаданных не только фамилии и имена авторов, приведенные в статье, но также варианты их написания на других языках. В результате работы соответствующего программного приложения был сгенерирован набор файлов в формате JATS, которые описывают каждую статью из обрабатываемого источника [33, 41].

Одной из структурных особенностей формата метаданных JATS является необходимость выбора основного языка представления статьи, а остальные языки объявляются альтернативными. Это создает сложности при формировании мультиязычных коллекций. Поэтому выбор основного языка представления – один из вопросов, которые приходится решать при создании xml-представления документов. Один из вариантов решения этой проблемы – использование языка, на котором написаны статьи, однако это не всегда позволяет организовать адекватный поиск в коллекциях цифровой библиотеки Lobachevskii-DML, потому что электронные коллекции этой библиотеки содержат в основном статьи на русском языке, а большая часть материалов ретро-коллекций – документы на дореформенном русском языке. При обработке таких документов возникают сложности в

написании названий статей и имен авторов, а также дополнительной информации, необходимой для формирования метаданных.

Сложности метаописания документов ретро-коллекций в формате JATS возникают и со статьями, опубликованными частями в различных номерах журнала, а также со статьями, которые имеют продолжения, причем об этом, как правило, говорится только в тексте статьи-продолжения, имеющей то же самое название.

2. АРХИВНЫЕ МАТЕМАТИЧЕСКИЕ ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ: ПРОБЛЕМЫ ФОРМИРОВАНИЯ МЕТАДАННЫХ

К архивным (ретро-коллекциям) мы относим электронные коллекции, которые содержат документы (статьи из периодических изданий (журналов), книги, препринты, сборники докладов конференций), напечатанные в период до широкого внедрения информационных технологий не только в процесс создания документа его авторами, но и в последующие этапы жизненного цикла этой публикации (см., например, [1, 3, 4]). Распространение научных знаний в этот период (обычно обозначаемый как «доцифровой») осуществлялось исключительно посредством печатных изданий. Как следствие, документы этих изданий, как правило, не содержат атрибутов, обязательных для изданий, распространяемых в Сети, таких, как предметные классификаторы, ключевые слова, аннотации, аффилиации авторов.

Отдельную категорию образуют исторические научные коллекции (ретро-коллекции), к которым можно отнести научные документы, опубликованные в периодических изданиях до начала XX века, а русскоязычные научные издания – до орфографической реформы 1918 года.

В работах [42–44] приведены результаты применения сервисов фабрики метаданных к документам ретро-коллекций цифровой библиотеки Lobachevskii-DML. Нормализация метаданных была проведена в них в соответствии с xml-схемами обязательного набора EuDML [29, 45].

Отметим наиболее важные проблемы формирования метаданных документов ретро-коллекций:

- разнообразие типов научных документов, размещенных в одном выпуске журнала, – статьи, доклады, письма, протоколы, объявления с отличающейся структурой оформления;

- отсутствие в статьях предметных классификаторов;
- отсутствие в статьях ключевых слов, характеризующих область исследования;
- отсутствие аннотаций к статьям;
- проблема с поиском в документе авторов статей – авторы могут быть указаны как в начале статьи, так и на последней ее странице;
- проблемы с полным описанием авторов: у автора статьи могут быть указаны только фамилия и начальная буква имени; встречаются сокращения фамилий авторов до инициалов (например, встречающееся сокращение «А. В.» соответствует статьям главного редактора А. Васильева (например, Известия. физ.-мат. общества, 1894 год, том 4));
- фамилия автора может снабжаться титулом (например, «Свящ. И. Максимовъ» (Известия физ.-мат. общества, 1915 год, том 11)), что требует использования отдельных шаблонов в методах экстракции метаданных;
- в статьях не указаны места работы авторов;
- в названиях теорем, а также в ссылках на статьи фамилии авторов приведены на языке оригинала;
- ссылки на литературу часто приведены в сносках либо непосредственно в тексте без полного библиографического описания;
- русскоязычные электронные ретро-коллекции содержат документы, использующие орфографию до реформы русского языка 1918 года.

3. КАКИЕ ЗНАНИЯ МОЖНО ИЗВЛЕЧЬ ИЗ WIKIDATA

Wikidata является базой знаний Википедии и центральной платформой управления данными для Википедии, а также родственных ей проектов (sister project) (см., например, [46, 47]). С момента запуска Wikidata в 2012 году на сайте этого проекта при участии более 5 млн. зарегистрированных пользователей собраны данные о 96633609 записях (информация на ноябрь 2021 года) (текущую статистику можно получить в [48]). Значительный профессиональный интерес к этому проекту объясняется тем, что Wikidata охватывает широкий спектр общих и специализированных знаний, актуальных во многих областях применения. Большая часть утверждений Wikidata снабжена сведениями об их происхождении, а

также дополнительными контекстными данными, такими как временная достоверность. Кроме того, приводимые данные связаны с внешними наборами данных во многих областях знаний, и информация продублирована на различных языках.

Объекты реального мира представлены в Wikidata элементами (items). Каждому элементу назначен числовой идентификатор с префиксом “Q”. Элементам соответствуют Wikipage in the Wikidata main namespace. Wikipage каждого элемента организована в виде свойств (properties) и утверждений (statements). Экземпляры свойств и утверждений также называют сущностями (entity), они имеют свои идентификаторы (с префиксом “Q” для утверждений и с префиксом “P” для свойств), которые служат важным источником метаданных item’a [49]. И у элементов, и у свойств имеются метка, описание и (многоязычные) псевдонимы.

Модель данных Wikidata приведена в [50]. Особенности работы с именованными сущностями в Wikidata выделены в работе [51]. Формулы присутствуют во всех математических статьях. Методы представления математических формул в Wikidata описаны в [52].

На страницах Wikidata, приведенных на рис. 1, представлена информация, которая использовалась при пополнении метаданных к статье А. Маркова «Распространение закона больших чисел на величины, зависящие друг от друга», опубликованной в номере 4 за 1906 год «Известий …». Рабочие процессы создания ретро-коллекции журнала «Известия …» и особенности формирования метаданных ее документов описаны в [43, 44]. Наиболее существенной является проблема идентификации авторов статей при фильтрации результатов запросов. Авторы статей этой коллекции указаны в выпусках журнала только фамилией и инициалами, иногда даже с одним инициалом (например, «А. Марковъ»). Поэтому при обработке результатов запросов потребовались фильтрация по нескольким признаков и дальнейшая проверка с привлечением экспертов.

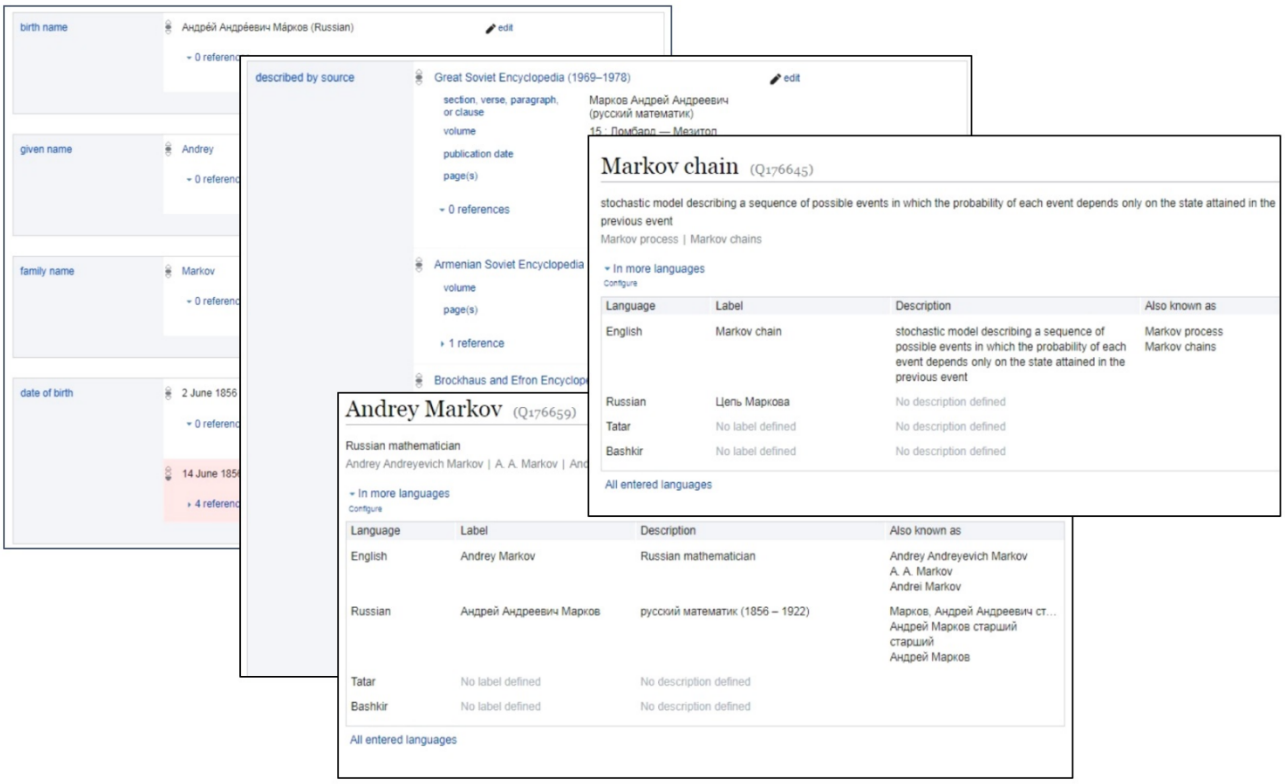

Рис. 1. Страницы Wikidata по запросам «Андрей Марков» и «Цепь Маркова».

Укажем основные свойства, значения которых были включены в состав метаданных: birth name (“Андре́й Андре́евич Ма́рков (Russian)”), given name (“Andrey”), family name (“Markov”), date of birth (“2 June 1856[Julian]”, “14 June 1856[Gregorian]”), date of death (“20 July 1922[Gregorian]”), occupation (“mathematician”, “statistician”, “university teacher”), field of work (“probability theory”, “mathematical analysis”, “number theory”), employer (“Saint Petersburg Academy of Sciences”, “Saint Petersburg State University”). Идентификаторы ID (“Q176659” и “Q176645”) также сохранены в метаданных – в дальнейшем с их помощью можно отслеживать обновление информации на соответствующих страницах Wikidata.

4. АЛГОРИТМЫ ПОПОЛНЕНИЯ МЕТАДАННЫХ ИНФОРМАЦИЕЙ ИЗ WIKIDATA

В данном разделе приведены алгоритмы формирования фундаментального набора метаданных документов ретро-коллекций цифровой библиотеки Lobachevskii-DML. Информация, которая по ряду причин оказалась недоступной для извлечения методами текстового и структурного анализа, была пополнена из открытых научных ресурсов Сети через систему поисковых запросов.

С помощью инструментов фабрики метаданных цифровой библиотеки Lobachevskii-DML была произведена обработка оцифрованных документов ретроколлекций. В результате удалось выделить из текстов следующие метаданные (в скобках приведены названия групп тегов в соответствии с xml-схемами JATS):

- название издания (“journal-title-group” и “trans-title-group”),
- время и место публикации (“publisher”, “pub-date”),
- название статьи (“title-group”) на одном из языков (дореформенном русском, английском, немецком, французском),
- фамилия автора с инициалами или только с одним инициалом (“contrib-group”),
- том и номер выпуска периодического издания (“volume”, “issue”),
- номера первой и последней страниц публикации (“star-page”, “end-page”).

Обработка сносок, часто присутствующих в текстах статей, позволяет получить информацию о названии и авторах (или только об одном из возможных авторов) тех статей, на которые указывают ссылки.

Для получения дополнительной информации, в частности, формирования фундаментального набора метаданных по схеме EuDML, предложен следующий алгоритм.

Имеющиеся метаданные преобразуются в csv-формат. Фамилии авторов и названия статей дополняются вариантами их написания на современном русском языке (в случае использования в документе дореформенной орфографии), а также производится транслитерация. Все названное необходимо для формирования шаблонов поисковых запросов.

На следующем этапе формируются поисковые запросы, включающие шаблоны, полученные на предыдущем этапе.

Далее выполняются стандартные операции по обработке полученных данных (см., например, [53]).

С помощью сформированных поисковых запросов можно уточнить (или дополнить) информацию об именах и отчествах авторов, месте работы, годах их публикационной активности, а также добавить URL сайтов, содержащих биографии и другую информацию об авторах.

Отметим, что описанный подход оказался результативным только в тех случаях, когда авторы документов электронных коллекций отражены в сетевом научном пространстве.

Для уточнения, а также пополнения уже сформированных метаданных были использованы открытые научные ресурсы, в частности, Wikipedia, Wikidata, DBPedia и Freebase [13, 54]. Для поиска и извлечения информации из Сети был применен инструментарий пакетов wikipedia и pywikibot (см., например, [55–57]). Также была разработана система SPARQL-запросов к ресурсам Wikidata.

Приведем теперь алгоритм извлечения метаданных из открытых научных ресурсов Сети, с обработкой полученных данных (см. Алгоритм 1).

```
<article-meta>
  <article-id>14_1_1_0</article-id>
  <title-group>
    <article-title xmllang="fr">Rapport sur les travaux de M. Pieri</article-title>
  </title-group>
  <contrib-group>
    <contrib contrib-type="author">
      <name-alternatives>
        <name>
          <surname xmllang="fr">Peano</surname>
          <string-name xmllang="fr">G. Peano</string-name>
        </name>
      </name-alternatives>
    </contrib>
  </contrib-group>
  <volume>14</volume>
  <volume-series>2</volume-series>
  <pub-date>
    <year>1904</year>
  </pub-date>
  <issue>1</issue>
  <issue-part>1</issue-part>
</article-meta>
```

Рис. 2. Фрагмент набора метаданных статьи G. Peano “Rapport sur les travaux de M. Pieri”, опубликованной в журнале «Известия …» (серия 2, том 14, номер 1 за 1904 год). Метаданные сформированы по схеме NISO JATS с помощью инструментов фабрики метаданных цифровой библиотеки Lobachevskii-DML. Из текста документа экстрагированы: название статьи на французском языке, фамилия и инициал автора, номера начальной и финальной страниц документа.

На вход алгоритма подается набор

$$M=\{d\_1.xml,d\_2.xml,\ldots,d\_m.xml\},$$

состоящий из файлов с метаданными документов ретро-коллекции. На рис. 2

приведен фрагмент такого файла. Из рассматриваемой статьи удалось извлечь только ее название (“Rapport sur le travaux de M. Pieri”) и фамилию автора (G. Peano). Отметим, что автор статьи в приведенном примере – известный математик Джузеппе Пеано (1858–1932); в настоящее время в Сети имеется информация об этом математике, как и о многих других авторах формируемой электронной ретро-коллекции.

Как уже было сказано, полученные наборы метаданных являются неполными, так как в соответствии со схемами интегрирующих цифровых математических библиотек требуется существенно больший объем метаинформации о научных документах. В частности, полученных метаданных недостаточно для формирования фундаментального набора по xml-схемам цифровой математической библиотеки EuDML [29].

Алгоритм 1: Извлечение метаданных из открытых научных источников Сети

```
load M=[d1.xml, d2.xml,..., dm.xml]
for d in M:
  # Разбор XML-дерева:
  md=parse(d).getroot()
 # Найти группу тегов с данными об авторах
 # (схема NISO JATS):
 for authors in md.findall(“./conrib_group/
              [@content_type=’authors’]”):
   # Выбрать группу тегов для каждого автора:
   for author in md.findall(“./contrib/
                    [@contrib_type=’author’]”):
     # Найти тег идентифицирующий имя автора и его инициалы:
  name_author_in_paper=author.find(‘string-name’)
     # Перевести и транслитерировать имя и фамилию автора:
•    if language(name_author_in_paper) !=’ru’:
•     name_author_ru=translate_ru()
•    if language(name_author_in_paper)==’ru’:
•     name_author_en=transliterate()
•    if language(name_author_in_paper)==’ru-old’:
•     name_author_ru=translate_ru_old()
•     name_author_en=transliterate()
•      # Сформировать список шаблонов поисковых запросов:
•    patterns=pattern_list()
•      # Выбрать и соединиться с точкой доступа
• #    (Wikipedia, Wikidata, DBPedia):
•    point=point_connect()
```

```
    results=[]
      # Поиск с каждым из шаблонов:
    for p in patterns:
     result=search(p)
          # Обработка результатов:
     extracting(result)
     cleaning(result)
     similarity(result)
     results.append(result)
    end for
      # Запись новых метаданных в соответствии с XML-схемой:
    normalization(results)
  end for
end for
```

В качестве источника пополнения метаданных нами были использованы открытые ресурсы Сети. Программные инструменты фабрики метаданных цифровой математической библиотеки Lobachevskii-DML на основе текстового анализа документов электронных ретро-коллекций позволяют извлекать такие метаданные, как название статьи, библиографические ссылки, диапазоны страниц, фамилии авторов на языке оригинала (русском, дореформенном русском, немецком, французском или английском). В настоящее время в Сети об авторах большинства статей формируемой электронной коллекции имеются сведения, которые отсутствовали в самих статьях. Это сделало возможным извлечение из сетевых ресурсов недостающей информации об авторах статей, в частности, о вариантах написания на различных языках их фамилий, имен и отчеств, мест работы в момент написания статьи (см. рис. 3 и 4).

Отметим, что имеется ряд сервисов связывания данных, содержащих объекты математического знания. Большинство из них имеет точку подключения (SPARQL endpoint) [58].

```
<article-meta>
  <article-id>14_1_1_0</article-id>
  <title-group>
    <article-title xmllang="fr">Rapport sur les
    travaux de M. Pieri</article-title>
  </title-group>
  <contrib-group>
    <contrib contrib-type="author">
      <name-alternatives>
        <name>
          <surname xmllang="fr">Peano</surname>
          <surname xmllang="ru">Пеано</surname>
          <given-names xmllang="fr">G.</given-names>
          <given-names xmllang="it">Giuseppe
          </given-names>
          <given-names xmllang="ru">Джузеппе
          </given-names>
        </name>
        <string-name xmllang="fr">G. Peano
        </string-name>
        <string-name xmllang="it">Giuseppe Peano
        </string-name>
        <string-name xmllang="ru">Джузеппе Пеано
        </string-name>
      </name-alternatives>
      <aff xmllang="fr">University of Turin</aff>
    </contrib>
  </contrib-group>
  <volume>14</volume>
  <volume-series>2</volume-series>
  <pub-date>
    <year>1904</year>
  </pub-date>
  <issue>1</issue>
  <issue-part>1</issue-part>
</article-meta>
```

Рис. 3. Фрагмент фундаментального набора метаданных, сформированных по Алгоритму 1. Ранее сформированный набор (см. рис. 2) был дополнен информацией об авторе статьи.

E	F	G	H	I	J	K	L
автор	исходное	название	исходное название статьи	WikidataURI	MathN	ZbMATHAuthorID	OpenLibraryID
H. Poincare		Rapport sur les travaux de M. Hilbert		Q81082		poincare.henri	OL7476098A
P. Mansion		Rapport sur les travaux de M. Barbarin		null		mansion.paul	OL3775794A
C. A. Laisant		Rapport sur les travaux de M. Lemoine		Q25318		laisant.ch-a	OL2426857A
G. Peano		Rapport sur les travaux de M. Pieri		Q191029		peano.giuseppe	OL32329A

Рис. 4. Дополнительные метаданные, экстрагированные из сетевых источников. В частности, получены URL страниц, содержащих упоминание о рассматриваемом документе.

На рис. 5 приведен пример обработки результата, полученного по запросу

в Wikidata и в ходе последующей фильтрации по ряду признаков, ограничивающих результаты запроса принадлежностью к научной деятельности. По запросу, содержащему фамилию автора статьи, в Wikidata было обнаружено 229 элементов. После проведения процедуры уточнения выделен элемент с меткой Q16648192, содержащий информацию об авторе статьи.

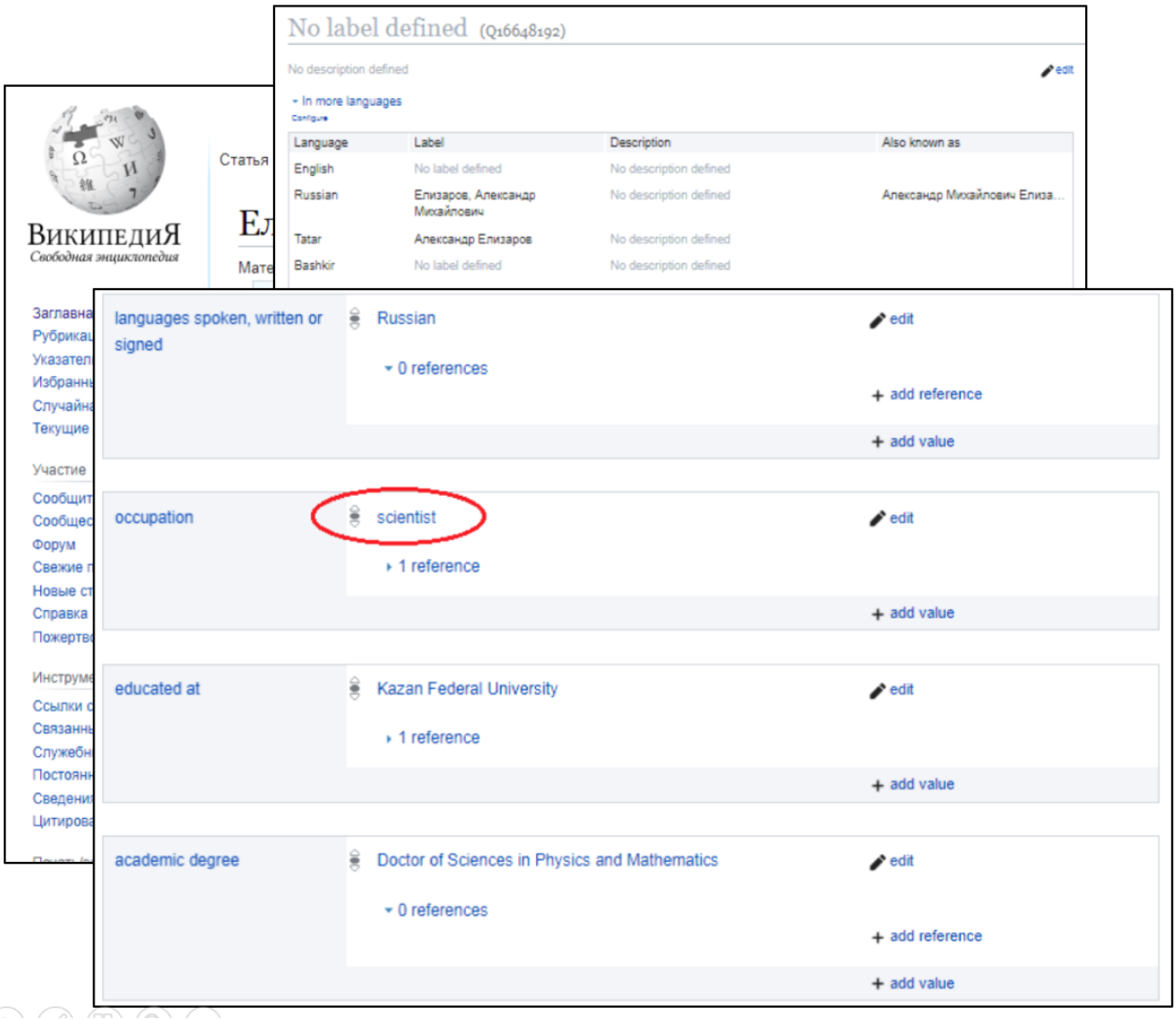

Рис. 5. Страница персоны в Wikidata. Отметим, что по данной персоне в Wikipedia имеется только русскоязычная страница, поэтому на странице Wikidata отсутствует label.

В таблице 1 приведены основные свойства (properties), информация из которых была использована при формировании дополнительных метаданных.

Таблица 1. Основные свойства Wikidata, используемые в алгоритме пополнения метаданных (на примере элемента с ID Q570859)

Property	ID	Пример	Jats_Tag
family name	P734	Chebotaryov	<surname>
given name	P735	Nikolai	<given-names>
name in native language	P1559	Николай Григорьевич Чеботарёв (Russian)	<string-name>
birth name	P1477	Николай Григорьевич Чеботарёв (Russian)	<string-name>
date of birth	P569	3 June 1894*Julian*, 15 June 1894*Gregorian*, 1894	<def-list>
date of death	P570	2 July 1947, 1947	<def-list>
occupation	P106	Mathematician(Q170790), university teacher (Q1622272)	<def-list>
employer	P108	Kazan Federal University (Q113788)	<aff>
member of	P463	Academy of Sciences of the USSR (Q2370801)	<aff-alterna-tives>
academic degree	P512	Doctor of Sciences in Physics and Mathematics (Q17281097)	<degrees>
field of work	P101	number theory (Q12479), algebra (Q3968), function theory (Q4455174)	<def-list>
notable work	P800	Chebotarev's density theorem (Q1425529), Chebotarev theorem on roots of unity (Q17007435)	<def-list>

Важным свойством, используемым в SPARQL-запросах к Wikidata, является свойство occupation (P106). С его помощью можно произвести фильтрацию результатов запроса, оставив только страницы с информацией о персонах, связанных с научной деятельностью (см. таблицу 2).

Таблица 2. Основные критерии отбора item по названию и количество соответствующих страниц

occupation (P106)	ID	Количество результатов
scientist	Q901	444354
mathematician	Q170790	35182
researcher	Q1650915	148544
university teacher	Q1622272	164512

Отметим, что при создании запросов к Wikidata учитываются синонимичные property, предоставляющие различные способы получения одних и тех же данных, например, свойства “name in native language” и “birth name” имеют одинаковые значения.

Ниже приведен Алгоритм 2 пополнения метаданных документов электронных коллекций с помощью запросов к Wikidata. На Рис. 6 приведены фрагменты класса Table и структуры item, используемые в алгоритме при формировании информации об авторе статьи.

```
class Table
{
    public List<item> familyname
    public List<item> initials
    public List<string> uri
    public List<item> Props
    public string type;
…..
}
```

```
struct item
{
    public string Property;
    public string ID;
    public string Jats_Tag;
    public string lang;
}
```

Рис. 6. Фрагменты класса и структуры, используемых для описания автора в Алгоритме 2.

Алгоритм 2: Пополнение метаданных документа цифровой коллекции

```
read metadata_set
List<string> authors_result = selected content from tag <authors>
#Список метаданных в xml формате
List <XElement> metadata
foreach authors_str in authors_results
        List <string> authors = Split(authors_str)
        foreach author in authors
                # поиск автора в Wikidata Wikidata,
                form SPARQL requests for Wikidata by family name (P734)
                # SPARQL запросы (Рис. 7, Рис. 8)
                get list Request_list from request
                filter by initials (from birth name or name in native language),
                filter by occupation set in Table 2
                if Request_list.Length>1 then необходима ручная экспертиза
                else
                List<Table> Props = new List<Table>
```

```
14:             fill in the attributes ID, Jats_Tag, Property for each class instance
15:             foreach Prop in Props
16:                 form SPARQL requests for Wikidata: property is Prop.ID
17:                 get content for Prop.Content
                #Формирование metadata set
18:             form a metadata_set using list Props
19:         metadata.Add(metadata_set)
20: form new metadata_set
21: save new metadata_set
```

Поиск происходит при помощи службы MediaWiki API [59]. Она позволяет вызывать MediaWiki API из SPARQL и получать результаты из запроса SPARQL. Ниже представлены некоторые запросы, которые используются в этом алгоритме.

Стандартный запрос поиска в Wikidata дополнительной информации по автору статьи (соответствует шагу 7 Алгоритма 2; осуществляет поиск в Wikidata страниц документов с фамилией автора статьи) представлен на рис. 7.

```
select ?item where {
?item rdfs:label "Елизаров"@ru.
?item wdt:P31 wd:Q101352.
}
```

Рис. 7. Запрос со свойством “instance of” (P31) с явным указанием сущности “family name” (Q101352).

Далее производится поиск по сущностям, полученным на шаге 7 Алгоритма 2. С помощью фильтрации по принадлежности к профессии (“scientist” или другого значения из Таблицы 2) результаты сужаются, в большинстве случаев до ссылок на страницы статей искомого автора (Рис. 8).

```
SELECT DISTINCT ?item ?itemLabel WHERE {
  SERVICE wikibase:label { bd:serviceParam wikibase:language "ru". }
  {
    SELECT DISTINCT ?item WHERE {
      ?item p:P734 ?statement0.
      ?statement0 (ps:P734/(wdt:P279*)) wd:Q21507140.
      ?item p:P106 ?statement1.
      ?statement1 (ps:P106/(wdt:P279*)) wd:Q901.
    }
    LIMIT 100
  }
}
```

Рис. 8. Поисковый запрос по сущности, полученной на предыдущем шаге алгоритма, с фильтрацией по значению “scientist” (Q901) свойства “occupation” (P106).

Запрос по получению всех метаданных, указанных в Таблице 1, представлен на рис. 9. Результат включает не только ссылку на сущность (entity), но и значение этой сущности.

```
select * where {
     wd:Q570859 wdt:P734 ?family_name_id.
  ?family_name_id rdfs:label ?family_name filter(lang(?family_name)
= 'ru')
  wd:Q570859 wdt:P735 ?given_name_id.
  ?given_name_id rdfs:label ?given_name filter(lang(?given_name) =
'ru')
  wd:Q570859 wdt:P1559 ?name_in_native_language.
  wd:Q570859 wdt:P1477 ?birth_name.
   wd:Q570859 wdt:P569 ?date_of_birth.
   wd:Q570859 wdt:P570 ?date_of_death.
   wd:Q570859 wdt:P106 ?occupation_id.
  ?occupation_id rdfs:label ?occupation filter(lang(?occupation) =
'ru')
   wd:Q570859 wdt:P108 ?employer_id.
  ?employer_id rdfs:label ?employer filter(lang(?employer) = 'ru')
  wd:Q570859 wdt:P463 ?member_of_id.
  ?member_of_id rdfs:label ?member_of filter(lang(?member_of) =
'ru')
   wd:Q570859 wdt:P512 ?academic_degree_id.
  ?academic_degree_id rdfs:label ?academic_degree filter(lang(?aca-
demic_degree) = 'ru')
```

```
   wd:Q570859 wdt:P101 ?field_of_work_id.
  ?field_of_work_id rdfs:label ?field_of_work fil-
ter(lang(?field_of_work) = 'ru')
   wd:Q570859 wdt:P800 ?notable_work_id.
  ?notable_work_id rdfs:label ?notable_work filter(lang(?nota-
ble_work) = 'ru')
}
```

Рис. 9. Запрос по получению всех метаданных, указанных в Таблице 1.

```
<contrib-group>
    <contrib contrib-type="author">
        <name-alternatives>
            <name>
                <surname id="Q21493235" xml:lang="ru">Чеботарёв</surname>
                <surname id="Q21493235" xml:lang="en">Chebotaryov</surname>
                <given-names id="Q5486169" xml:lang="ru">Николай</given-names>
                <given-names id="Q5486169" xml:lang="en">Nikolai</given-names>
                <string-name id="P1559" xml:lang="ru">Николай Григорьевич Чеботарёв</string-name
                <string-name id="P1477" xml:lang="ru">Николай Григорьевич Чеботарёв</string-name
            </name>
        </name-alternatives>
        <bio>
            <def-list id="P569">
                <def-item>3 June 1894 Julian</def-item>
                <def-item>15 June 1894 Gregorian, </def-item>
                <def-item>1894</def-item>
            <def-list>
            <def-list id="P570">
                <def-item>2 July 1947</def-item>
                <def-item>1947</def-item>
            <def-list>
            <def-list id="P106">
                <def-item id="Q170790">mathematician</def-item>
                <def-item id="Q1622272">university teacher</def-item>
            <def-list>
            <def-list id="P101">
                <def-item id="Q12479">number theory</def-item>
                <def-item id="Q3968">algebra</def-item>
                <def-item id="Q4455174">function theory</def-item>
            <def-list>
            <def-list id="P800">
                <def-item id="Q1425529">Chebotarev's density theorem</def-item>
                <def-item id="Q17007435">Chebotarev theorem on roots of unity</def-item>
            <def-list>
        </bio>
        <aff id="Q113788">Kazan Federal University</aff>
        <aff-alternatives id="Q2370801">Academy of Sciences of the USSR </aff-alternatives>
        <degrees id="Q17281097">Doctor of Sciences in Physics and Mathematics </degrees>
    </contrib>
</contrib-group>
```

Рис. 10. Фрагмент Jats-представления документа с метаописанием автора статьи по информации, полученной из Wikidata.

Далее производится обработка результатов SPARQL-запросов, включающая проведение трансформации в набор метаданных в формате JATS. Фрагмент полученных метаданных приведен на рис. 10.

Отметим, что при внутреннем представлении документов электронной коллекции используются id сущности Wikidata.

ЗАКЛЮЧЕНИЕ

Представлен метод создания обязательного набора метаданных документов электронных ретро-коллекций цифровой математической библиотеки Lobachevskii-DML. Названный набор соответствует известной схеме метаданных EuDML, широко используемой в настоящее время в цифровых математических библиотеках. При формировании такого набора метаданных документов, опубликованных в «доцифровой» период и включаемых в ретро-коллекции, возникает ряд проблем, связанных в первую очередь с недостаточностью имеющейся информации, необходимой для создания метаданных. Поэтому в качестве источника пополнения такой информации предложено использовать открытые ресурсы Веба, в частности, Wikidata.

С помощью программных инструментов созданной ранее фабрики метаданных цифровой математической библиотеки Lobachevskii-DML реализованы основные процессы текстового анализа документов электронных ретро-коллекций и выделены именованные сущности. Разработана система запросов для организации поиска в Сети информации, необходимой для получения метаданных, с последующей экстракцией информационных объектов. После автоматизированного проведения фильтрации и нормализации полученная информация включается в набор метаданных. Приведены алгоритмы пополнения метаданных документов ретро-коллекций информацией, полученной из Wikidata.

Одними из основных результатов проведенного исследования стали формирование обязательного набора метаданных ретро-коллекции статей журнала «Известия физико-математического общества при Казанском университете» и ее включение в состав цифровой математической библиотеки Lobachevskii-DML.

Благодарности

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

СПИСОК ЛИТЕРАТУРЫ

1. *Bartling S., Friesike S*. Towards Another Scientific Revolution // In: Bartling S., Friesike S. (Eds.) Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer International Publishing. 2014. P. 3–15 (2014). https://doi.org/10.1007/978-3-319-00026-8_1.

2. *Carette J., Farmer W.M., Kohlhase M., Rabe F.* Big Math and the One-Brain Barrier: The Tetrapod Model of Mathematical Knowledge // Math. Intelligencer. 2021. Vol. 43. P. 78–87 (2021). https://doi.org/10.1007/s00283-020-10006-0.

3. *Елизаров А.М., Зуев Д.С., Липачёв Е.К.* Управление жизненным циклом электронных публикаций в информационной системе научного журнала // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2014. № 4. С. 81–88.

4. *Binfield P.* Novel Scholarly Journal Concepts // In: Bartling S., Friesike S. (Eds.) Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer International Publishing, 2014. P. 155–163. https://doi.org/10.1007/978-3-319-00026-8_10.

5. *Ataeva O., Kalenov N., Serebriakov V., Sotnikov A*. Informational Infrastructure of the Common Digital Space of Scientific Knowledge // CEUR Workshop Proceedings. 2021. Vol. 2990. P. 1–10. URL: http://ceur-ws.org/Vol-2990/rpaper1.pdf, last accessed 2021/11/07.

6. *Ion P.D.F.* Mathematics and the World Wide Web // In: Carette J., Aspinall D., Lange C., Sojka P., Windsteiger W. (Eds.) Intelligent Computer Mathematics. CICM 2013. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. 2013. Vol. 7961. https://doi.org/10.1007/978-3-642-39320-4_15.

7. *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. Vol. 10383. P. 56–69. https://doi.org/10.1007/978-3-319-62075-6_5.

8. Developing a 21st Century Global Library for Mathematics Research.

Washington: The National Academies Press, 2014. 142 p. https://doi.org/10.17226/18619.

9. *Xie I., Matusiak K.* Discover Digital Libraries: Theory and Practice. Elsevier Inc., 2016.

10. Born-digital. URL: https://en.wikipedia.org/wiki/Born-digital, last accessed 2021/11/07.

11. Author Guide – ScholarOne Manuscripts. Clarivate Analytics. 2019. P. 1–70. URL: https://clarivate.com/webofsciencegroup/wp-content/uploads/sites/2/dlm_uploads/2019/10/ScholarOne-Manuscripts-Author-Guide.pdf, last accessed 2021/11/07.

12. Author tutorials. Writing a journal manuscript. Springer Nature Switzerland AG, 2021. URL: https://www.springernature.com/gp/authors/campaigns/writing-a-manuscript, last accessed 2021/11/07.

13. *Gafurova P., Elizarov A., Lipachev E.* Algorithms for Integration of Unstructured Mathematical Documents into the Common Digital Space of Scientific Knowledge // CEUR Workshop Proceedings.2021. Vol. 2990. P. 39–49. URL: http://ceur-ws.org/Vol-2990/rpaper4.pdf, last accessed 2021/11/07.

14. *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.

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

16. *Elizarov A.M., Lipachev E.K., Khaydarov S.M.* Automated system of services for processing of large collections of scientific documents // CEUR Workshop Proceedings. 2016. Vol. 1752. P. 58–64.

17. *Tkaczyk D.* New Methods for Metadata Extraction from Scientific Literature. arXiv:1710.10201v1. 2017. URL: https://arxiv.org/pdf/1710.10201v1.pdf, last accessed 2021/09/09.

18. Universal Decimal Classification. URL: https://udcc.org/index.php, last accessed 2021/09/09.

19. MSC2020–Mathematics Subject Classification System.
URL: https://mathscinet.ams.org/msnhtml/msc2020.pdf, last accessed 2021/09/09.

20. *Řehůřek R., Sojka P.* Automated Classification and Categorization of Mathematical Knowledge // In: Autexier S., Campbell J., Rubio J., Sorge V., Suzuki M., Wiedijk F. (Eds.) Intelligent Computer Mathematics. CICM 2008. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. 2008. Vol. 5144. P. 543–557. https://doi.org/10.1007/978-3-540-85110-3_44.

21. *Хайдаров Ш.М., Ямалутдинова Г.Ш.* Рекомендательная система классификации физико-математических документов // Научный сервис в сети Интернет: труды XX Всероссийской научной конференции (17–22 сентября 2018 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2018. С. 480–486.
URL: https://doi.org/ 10.20948/abrau-2018-57. http://keldysh.ru/abrau/2018/theses/57.pdf.

22. *Schubotz M., Scharpf P., Teschke O., Kühnemund A., Breitinger C., Gipp B*. AutoMSC: Automatic Assignment of Mathematics Subject Classification Labels // In: Proceedings of the 13th Conference on Intelligent Computer Mathematics. 2020. arXiV:2005.12099v1. 25 May 2020.

23. *Nevzorova O., Almukhametov D.* Towards a Recommender System for the Choice of UDC Code for Mathematical Articles // CEUR Workshop Proceedings. 2021. Vol. 3036. P. 54–62.
URL: http://ceur-ws.org/Vol-3036/paper04.pdf, last accessed 2021/11/07.

24. *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.

25. *Elizarov A.M., Lipachev E.K., Zuev D.S*. Digital Mathematical Libraries: Overview of Implementations and Content Management Services // CEUR Workshop Proceedings. 2017. Vol. 2022. P. 317–325.

26. *Elizarov A.M., Lipachev E.K*. Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // CEUR Workshop Proceedings. 2017. Vol. 2022. P. 326–333. URL: http://ceur-ws.org/Vol-2022/paper50.pdf, last accessed 2021/11/07.

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

URL: http://ceur-ws.org/Vol-2523/invited08.pdf, last accessed 2021/11/21.

28. *Гафурова П.О., Елизаров А.М., Липачёв Е.К.* Базовые сервисы фабрики метаданных 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.

29. EuDML metadata schema specification (v2.0–final). https://initiative.eudml.org/eudml-metadata-schema-specification-v20-final, last accessed 2021/11/11.

30. *Elizarov A., Lipachev E.* Digital Libraries and the Common Digital Space of Mathematical Knowledge // CEUR Workshop Proceedings. 2021. Vol. 2990. P. 25–38. URL: http://ceur-ws.org/Vol-2990/rpaper3.pdf, last accessed 2021/11/07.

31. Электронная коллекция: Труды математического центра им. Н. И. Лобачевского. URL: https://lobachevskii-dml.ru/journal/tmt, last accessed 2021/11/07.

32. Электронная коллекция: «Известия физико-математического общества при Казанском университете». URL: https://lobachevskii-dml.ru/journal/izfmo2, https://lobachevskii-dml.ru/journal/izfmo3, last accessed 2021/11/07.

33. *Elizarov A., Lipachev E.* Digital Library Metadata Factories // CEUR Workshop Proceedings. 2021. Vol. 2813. P. 13–21. URL: http://ceur-ws.org/Vol-2813/rpaper01.pdf, last accessed 2021/11/07.

34. *Elizarov A.M., Khaydarov Sh.M., Lipachev E.K.* Scientific Documents Ontologies for Semantic Representation of Digital Libraries // In: Proceedings of the 2nd Russia and Pacific Conference on Computer Technology and Applications (RPC 2017). IEEE. 2017. P. 1–5. https://doi.org/10.1109/RPC.2017.8168064.

35. *Elizarov A., Lipachev E.* Methods of Processing Large Collections of Scientific Documents and the Formation of Digital Mathematical Library // CEUR Workshop Proceedings. 2020. V. 2543. P. 354–360. URL: http://ceur-ws.org/Vol-2543/spaper05.pdf, last accessed 2021/11/07.

36. *Lane H., Hapke H., Howard C.* Natural Language Processing in Action: Under-standing, analyzing, and generating text with Python. Manning Publications, 2019.

37. Natasha. URL: https://github.com/natasha/natasha, last accessed

2021/11/07.

38. Проект Natasha. Набор качественных открытых инструментов для обработки естественного русского языка (NLP).
URL: https://habr.com/ru/post/516098/, last accessed 2021/11/07.

39. *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. 99–10.
URL: https://www.emis.de/proceedings/TIEP2013/07bouche_rakosnik.pdf, last accessed 2021/11/11.

40. Journal Article Tag Suite. URL: https://jats.nlm.nih.gov/about.html, last accessed 2021/01/05.

41. *Gafurova P.O., Elizarov A.M., Lipachev E.K., Khammatova D.M*. Metadata Normalization Methods in the Digital Mathematical Library // CEUR Workshop Proceedings. 2020. Vol. 2543. P. 136–148.
URL: http://ceur-ws.org/Vol-2543/rpaper13.pdf, last accessed 2021/11/07.

42. *Гафурова П.О., Елизаров А.М., Липачёв Е.К.* Lobachevskii-DML: формирование архивных математических коллекций // Научный сервис в сети Интернет: труды XXII Всероссийской научной конференции. М.: ИПМ им. М.В. Келдыша, 2020. С. 171–183. https://doi.org/10.20948/abrau-2020-23.

43. *Gafurova P.O., Elizarov A.M., Lipachev E.K.* Metadata Extraction Methods for Organizing a Retro-Collection in the Lobachevskii Digital Mathematical Library // CEUR Workshop Proceedings. 2020. Vol. 2784. P. 62–71.
URL: http://ceur-ws.org/Vol-2784/rpaper06.pdf, last accessed 2021/11/07.

44. *Гафурова П.О., Елизаров А.М., Липачёв Е.К.* Алгоритмы формирования метаданных математических ретро-коллекций на основе анализа структурных особенностей документов // Электронные библиотеки. 2021. Т. 24, №2. С. 238–271. https://doi.org/10.26907/1562-5419-2021-24-2-238-270.

45. *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 2021/11/11.

46. *Vrandečić D., Krötzsch M.* Wikidata: a free collaborative knowledgebase // Communications of the ACM. October 2014. Vol. 57. Issue 10. P. 78–85.
https://doi.org/10.1145/2629489.

47. Wikipedia: Wikidata. URL: https://en.wikipedia.org/wiki/Wikidata, last accessed 2021/11/07.

48. Statistics – Wikidata.
URL: https://www.Wikidata.org/wiki/Special:Statistics, last accessed 2021/11/07.

49. Wikidata: Glossary.
URL: https://www.Wikidata.org/wiki/Wikidata:Glossary, last accessed 2021/11/07.

50. *Erxleben F., Günther M., Krötzsch M., Mendez J., Vrandečić D*. Introducing Wikidata to the Linked Data Web // In: Mika P. et al. (Eds.) The Semantic Web – ISWC 2014. ISWC 2014. Lecture Notes in Computer Science. Springer, Cham. 2014. Vol. 8796. P. 50–65. https://doi.org/10.1007/978-3-319-11964-9_4.

51. *Geiß J., Spitz A., Gertz M.* NECKAr: A Named Entity Classifier for Wikidata // In: Rehm G., Declerck T. (Eds.) Language Technologies for the Challenges of the Digital Age. GSCL 2017. Lecture Notes in Computer Science. Springer, Cham. 2018. Vol. 10713. P. 115–129. https://doi.org/10.1007/978-3-319-73706-5_10.

52. *Scharpf Ph., Schubotz M., Gipp B.* Mathematics in Wikidata // CEUR Workshop Proceedings. 2021. Vol. 2982. P. 1–14.
URL: http://ceur-ws.org/Vol-2982/paper-1.pdf, last accessed 2021/11/07.

53. *Knoblock C.A., Szekely P.* A scalable architecture for extracting, aligning, link-ing, and visualizing multi-Int data // Proc. SPIE 9499, Next-Generation Analyst III, 949907 (15 May 2015). https://doi.org/10.1117/12.2177119.

54. *Андреичев М.Д., Гафурова П.О., Елизаров А.М., Липачёв Е.К.* Пополнение метаданных документов математических цифровых ретро-коллекций методом семантических сетей // Научный сервис в сети Интернет: труды XXIII Всероссийской научной конференции (20–23 сентября 2021 г., онлайн). М.: ИПМ им. М.В. Келдыша, 2021. С. 22–33. https://doi.org/10.20948/abrau-2021-22. URL: https://keldysh.ru/abrau/2021/theses/22.pdf, last accessed 2021/11/07.

55. *Ayers P., Matthews C., Yates B.* How Wikipedia Works: And How You Can Be a Part of It. No Starch Press, San Francisco, CA, 2008.

56. Wikipedia Documentation.
URL: https://wikipedia.readthedocs.io/en/latest/code.html, last accessed 2021/11/07.

57. Pywikibot Documentation.
URL: https://doc.wikimedia.org/pywikibot/master/index.html, last accessed 2021/11/07.

58. SPARQL Query Language for RDF/W3C.
URL: https://www.w3.org/TR/rdf-sparql-query/. last accessed 2021/11/07.

59. MediaWiki is a collaboration and documentation platform brought to you by a vibrant community. URL: https://www.mediawiki.org/wiki/MediaWiki, last accessed 2021/11/07.

EXTRACTION OF WIKIDATA KNOWLEDGE FOR THE METADATA FORMATION FOR DOCUMENTS OF ELECTRONIC MATHEMATICAL COLLECTIONS

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] *Institute of Information Technologies and Intelligent Systems, Kazan (Volga Region) Federal University, ul. Kremlyovskaya, 35, Kazan, 420008*

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

Abstract

Methods for creating digital mathematical collections that include unstructured sets of documents are presented. These sets contain materials from scientific conferences, as well as articles from the archives of mathematical journals of the "pre-digital" period.

Using the software tools of the metadata factory of the digital mathematical library Lobachevskii DML, a mandatory set of metadata for digital collection documents was formed. To refine and replenish the metadata sets, knowledge extraction methods from Wikidata were used.

To search Wikidata for information about digital collection documents and their authors, a system of SPARQL queries has been developed. A set of Wikidata entities is defined, which determine the features of the search, as well as the subsequent filtering of the results.

Methods for clarifying and supplementing the bibliographic references given in the articles are proposed. When forming the metadata of documents of retrocollec-

tions, a search was made in Wikidata for information about the years of life of the authors of articles, as well as URLs of web pages with information about articles and their authors. The results of the formation of several new digital collections of the Lobachevskii-DML digital library are presented.

***Keywords**: Wikidata, metadata, metadata factory, digital mathematical collection, retrodigitized mathematical collection, Digital Mathematical Libraries, Lobachevskii-DML.*

REFERENCES

1. *Bartling S., Friesike S*. Towards Another Scientific Revolution // In: Bartling S., Friesike S. (Eds.) Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer International Publishing. 2014. P. 3–15. https://doi.org/10.1007/978-3-319-00026-8_1.

2. *Carette J., Farmer W.M., Kohlhase M., Rabe F.* Big Math and the One-Brain Barrier: The Tetrapod Model of Mathematical Knowledge // Math. Intelligencer. 2021. Vol. 43. P. 78–87. https://doi.org/10.1007/s00283-020-10006-0.

3. *Elizarov A.M., Zuev D.S., Lipachev E.K*. Lifecycle Management of Electronic Publications in Information Systems Scientific Journal // Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies. 2014. No. 4. P. 81-88.

4. *Binfield P.* Novel Scholarly Journal Concepts // In: Bartling S., Friesike S. (Eds.) Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer International Publishing, 2014. P. 155–163. https://doi.org/10.1007/978-3-319-00026-8_10.

5. *Ataeva O., Kalenov N., Serebriakov V., Sotnikov A*. Informational Infrastructure of the Common Digital Space of Scientific Knowledge // CEUR Workshop Proceedings 2990 (2021) 1–10. URL: http://ceur-ws.org/Vol-2990/rpaper1.pdf, last accessed 2021/11/07.

6. *Ion P.D.F.* Mathematics and the World Wide Web // In: Carette J., Aspinall D., Lange C., Sojka P., Windsteiger W. (Eds.) Intelligent Computer Mathematics. CICM 2013. Lecture Notes in Computer Science. 2013. Vol. 7961. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-39320-4_15.

7. *Ion P.D.F., Watt S.M.* The Global Digital Mathematics Library and the

International Mathematical Knowledge Trust // ICM 2017: Intelligent Computer Mathematics. Lecture Notes in Artificial Intelligence. 2017. Vol. 10383. P. 56–69. https://doi.org/10.1007/978-3-319-62075-6_5.

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

9. *Xie I., Matusiak K.* Discover Digital Libraries: Theory and Practice. Elsevier Inc., 2016.

10. Born-digital. URL: https://en.wikipedia.org/wiki/Born-digital, last accessed 2021/11/07.

11. Author Guide – ScholarOne Manuscripts. Clarivate Analytics. 2019. P. 1–70. URL: https://clarivate.com/webofsciencegroup/wp-content/uploads/sites/2/dlm_uploads/2019/10/ScholarOne-Manuscripts-Author-Guide.pdf, last accessed 2021/11/07.

12. Author tutorials. Writing a journal manuscript. Springer Nature Switzerland AG, 2021. URL: https://www.springernature.com/gp/authors/campaigns/writing-a-manuscript, last accessed 2021/11/07.

13. *Gafurova P., Elizarov A., Lipachev E.* Algorithms for Integration of Unstructured Mathematical Documents into the Common Digital Space of Scientific Knowledge // CEUR Workshop Proceedings. 2021. Vol. 2990. P. 39–49. URL: http://ceur-ws.org/Vol-2990/rpaper4.pdf, last accessed 2021/11/07.

14. *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. Vol. 48, No. 2. P. 81–85. https://doi.org/10.3103/S000510551402006X.

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

16. *Elizarov A.M., Lipachev E.K., Khaydarov S.M.* Automated system of services for processing of large collections of scientific documents // CEUR Workshop Proceedings. 2016. Vol. 1752. P. 58–64.

17. *Tkaczyk D.* New Methods for Metadata Extraction from Scientific

Literature. arXiv:1710.10201v1. 2017. URL: https://arxiv.org/pdf/1710.10201v1.pdf, last accessed 2021/09/09.

18. Universal Decimal Classification. URL: https://udcc.org/index.php, last accessed 2021/09/09.

19. MSC2020–Mathematics Subject Classification System.
URL: https://mathscinet.ams.org/msnhtml/msc2020.pdf, last accessed 2021/09/09.

20. *Řehůřek R., Sojka P.* Automated Classification and Categorization of Mathematical Knowledge // In: Autexier S., Campbell J., Rubio J., Sorge V., Suzuki M., Wiedijk F. (Eds.) Intelligent Computer Mathematics. CICM 2008. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. 2008. Vol. 5144. P. 543–557.
https://doi.org/10.1007/978-3-540-85110-3_44.

21. *Khaydarov S.M., Yamalutdinova G.S.* Recommender System of Physical and Mathematical Documents Classification. CEUR Workshop Proceedings. 2018. Vol. 2260. P. 480–486. URL: http://ceur-ws.org/Vol-2260/57_480-486.pdf, last accessed 2021/11/07.

22. *Schubotz M., Scharpf P., Teschke O., Kühnemund A., Breitinger C., Gipp B*. AutoMSC: Automatic Assignment of Mathematics Subject Classification Labels // In: Proceedings of the 13th Conference on Intelligent Computer Mathematics. 2020. arXiV:2005.12099v1. 25 May 2020.

23. *Nevzorova O., Almukhametov D.* Towards a Recommender System for the Choice of UDC Code for Mathematical Articles // CEUR Workshop Proceedings. 2021. Vol. 3036. P. 54–62.
URL: http://ceur-ws.org/Vol-3036/paper04.pdf, last accessed 2021/11/07.

24. *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.

25. *Elizarov A.M., Lipachev E.K., Zuev D.S*. Digital Mathematical Libraries: Overview of Implementations and Content Management Services // CEUR Workshop Proceedings. 2017. Vol. 2022. P. 317–325.

26. *Elizarov A.M., Lipachev E.K*. Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // CEUR Workshop Proceedings. 2017. Vol. 2022. P. 326–333. URL: http://ceur-ws.org/Vol-2022/paper50.pdf, last accessed 2021/11/07.

27. *Elizarov A.M., Lipachev E.K.* Big Math Methods in Lobachevskii-DML Digital Library // CEUR Workshop Proceedings. 2019. Vol. 2523. P. 59–72.
URL: http://ceur-ws.org/Vol-2523/invited08.pdf, last accessed 2021/11/21.

28. *Gafurova P.O., Elizarov A.M., Lipachev E.K.* Basic Services of Factory Metadata Digital Mathematical Library Lobachevskii-DML // Russian Digital Libraries Journal. 2020. V. 23, No. 3. P. 336–381.
https://doi.org/10.26907/1562-5419-2020-23-3-336-381.

29. EuDML metadata schema specification (v2.0–final).
https://initiative.eudml.org/eudml-metadata-schema-specification-v20-final, last accessed 2021/11/11.

30. *Elizarov A., Lipachev E.* Digital Libraries and the Common Digital Space of Mathematical Knowledge // CEUR Workshop Proceedings. 2021. Vol. 2990. P. 25–38.
URL: http://ceur-ws.org/Vol-2990/rpaper3.pdf, last accessed 2021/11/07.

31. Digital Collection: Proceedings of Lobachevskii mathematical center.
URL: https://lobachevskii-dml.ru/journal/tmt, last accessed 2021/11/07

32. Digital Collection: “Izvestia of the Physics and Mathematics Society at Kazan University” (“Bulletin de la Société Physico-Mathémaique de Kasan”).
URL: https://lobachevskii-dml.ru/journal/izfmo2,
https://lobachevskii-dml.ru/journal/izfmo3, last accessed 2021/11/07.

33. *Elizarov A., Lipachev E.* Digital Library Metadata Factories // CEUR Workshop Proceedings. 2021. Vol. 2813. P. 13–21.
URL: http://ceur-ws.org/Vol-2813/rpaper01.pdf, last accessed 2021/11/07.

34. *Elizarov A.M., Khaydarov Sh.M., Lipachev E.K.* Scientific Documents Ontologies for Semantic Representation of Digital Libraries // In: Proceedings of the 2nd Russia and Pacific Conference on Computer Technology and Applications (RPC 2017). IEEE. 2017. P. 1–5. https://doi.org/10.1109/RPC.2017.8168064.

35. *Elizarov A., Lipachev E.* Methods of Processing Large Collections of Scientific Documents and the Formation of Digital Mathematical Library // CEUR Workshop Proceedings. 2020. V. 2543. P. 354–360.
URL: http://ceur-ws.org/Vol-2543/spaper05.pdf, last accessed 2021/11/07.

36. *Lane H., Hapke H., Howard C.* Natural Language Processing in Action: Under-standing, analyzing, and generating text with Python. Manning Publications, 2019.

37. Natasha. URL: https://github.com/natasha/natasha, last accessed 2021/11/07.

38. Proekt Natasha. Nabor kachestvennyh otkrytyh instrumentov dlya obrabotki estestvennogo russkogo yazyka (NLP).
URL: https://habr.com/ru/post/516098/, last accessed 2021/11/07.

39. *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. 99–10.
URL: https://www.emis.de/proceedings/TIEP2013/07bouche_rakosnik.pdf, last accessed 2021/11/11.

40. Journal Article Tag Suite. URL: https://jats.nlm.nih.gov/about.html, last accessed 2021/01/05.

41. *Gafurova P.O., Elizarov A.M., Lipachev E.K., Khammatova D.M*. Metadata Normalization Methods in the Digital Mathematical Library // CEUR Workshop Proceedings. 2020. Vol. 2543. P. 136–148.
URL: http://ceur-ws.org/Vol-2543/rpaper13.pdf, last accessed 2021/11/07.

42. *Гафурова П.О., Gafurova P. O., Elizarov A. M., Lipachev E. K*. Lobachevskii-DML: Formation of Archival Mathematical Collections // Nauchnyj servis v seti Internet: trudy XXII Vserossijskoj nauchnoj konferencii. M.: IPM im. M.V. Keldysha, 2020. S. 171–183. https://doi.org/10.20948/abrau-2020-23.

43. *Gafurova P.O., Elizarov A.M., Lipachev E.K.* Metadata Extraction Methods for Organizing a Retro-Collection in the Lobachevskii Digital Mathematical Library // CEUR Workshop Proceedings. 2020. Vol. 2784. P. 62–71.
URL: http://ceur-ws.org/Vol-2784/rpaper06.pdf, last accessed 2021/11/07.

44. *Gafurova P.O., Elizarov A.M., Lipachev E.K.* Algorithms for Formation of Metadata Mathematical Retro Collections Based on Analysis of Structural Features of Documents // Russian Digital Libraries Journal. 2021. Vol. 24. No. 2. P. 238–271.
https://doi.org/10.26907/1562-5419-2021-24-2-238-270.

45. *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 2021/11/11.

46. *Vrandečić D., Krötzsch M.* Wikidata: a free collaborative knowledgebase // Communications of the ACM. October 2014. Vol. 57. Issue 10. P. 78–85.

https://doi.org/10.1145/2629489.

47. Wikipedia: Wikidata. URL: https://en.wikipedia.org/wiki/Wikidata, last accessed 2021/11/07.

48. Statistics – Wikidata.
URL: https://www.Wikidata.org/wiki/Special:Statistics, last accessed 2021/11/07.

49. Wikidata: Glossary.
URL: https://www.Wikidata.org/wiki/Wikidata:Glossary, last accessed 2021/11/07.

50. *Erxleben F., Günther M., Krötzsch M., Mendez J., Vrandečić D*. Introducing Wikidata to the Linked Data Web // In: Mika P. et al. (Eds.) The Semantic Web – ISWC 2014. ISWC 2014. Lecture Notes in Computer Science. Springer, Cham. 2014. Vol. 8796. P. 50–65. https://doi.org/10.1007/978-3-319-11964-9_4.

51. *Geiß J., Spitz A., Gertz M.* NECKAr: A Named Entity Classifier for Wikidata // In: Rehm G., Declerck T. (Eds.) Language Technologies for the Challenges of the Digital Age. GSCL 2017. Lecture Notes in Computer Science. Springer, Cham. 2018. Vol. 10713. P. 115–129. https://doi.org/10.1007/978-3-319-73706-5_10.

52. *Scharpf Ph., Schubotz M., Gipp B.* Mathematics in Wikidata // CEUR Workshop Proceedings. 2021. Vol. 2982. P. 1–14.
URL: http://ceur-ws.org/Vol-2982/paper-1.pdf, last accessed 2021/11/07.

53. *Knoblock C.A., Szekely P.* A scalable architecture for extracting, aligning, link-ing, and visualizing multi-Int data // Proc. SPIE 9499, Next-Generation Analyst III, 949907 (15 May 2015). https://doi.org/10.1117/12.2177119.

54. *Andreichev M.D., Gafurova P.O., Elizarov A.M., Lipachev E.K.* Replenishment of Documents of Mathematical Digital Retro-collections by Searching in Semantic Web. Nauchnyj servis v seti Internet: trudy XXIII Vserossijskoj nauchnoj konferencii (20–23 sentyabrya 2021 g., onlajn). M.: IPM im. M.V. Keldysha, 2021. S. 22–33. https://doi.org/10.20948/abrau-2021-22.
URL: https://keldysh.ru/abrau/2021/theses/22.pdf, last accessed 2021/11/07.

55. *Ayers P., Matthews C., Yates B.* How Wikipedia Works: And How You Can Be a Part of It. No Starch Press, San Francisco, CA, 2008.

56. Wikipedia Documentation.
URL: https://wikipedia.readthedocs.io/en/latest/code.html, last accessed 2021/11/07.

57. Pywikibot Documentation.
URL: https://doc.wikimedia.org/pywikibot/master/index.html,

last accessed 2021/11/07.

58. SPARQL Query Language for RDF/W3C.
URL: https://www.w3.org/TR/rdf-sparql-query/. last accessed 2021/11/07.

59. MediaWiki is a collaboration and documentation platform brought to you by a vibrant community. URL: https://www.mediawiki.org/wiki/MediaWiki, last accessed 2021/11/07.

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 Michailovich 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, 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.

URL: https://kpfu.ru/Evgeny.Lipachev.

email: elipachev@gmail.com;

ORCID: 0000-0001-7789-2332

Материал поступил в редакцию 11 ноября 2021 года