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).

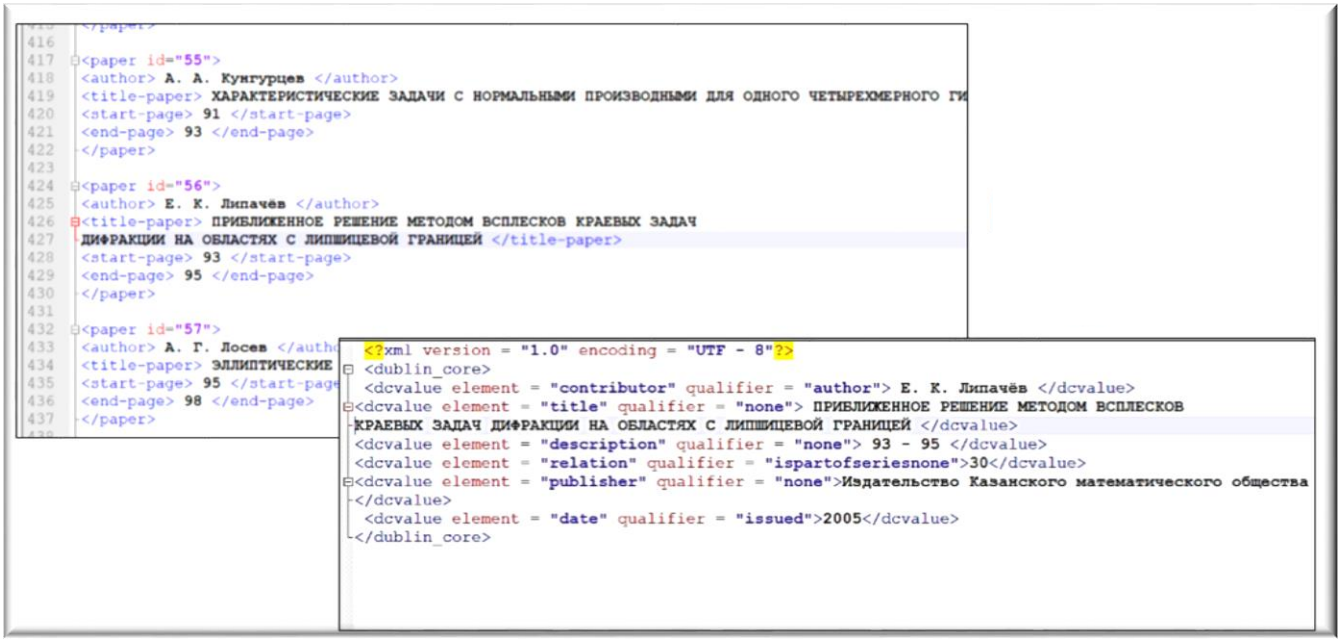


Рис. 1. Преобразование метаданных в спецификацию Dublin Core с учётом специфики цифрового хранилища DSpace.

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.

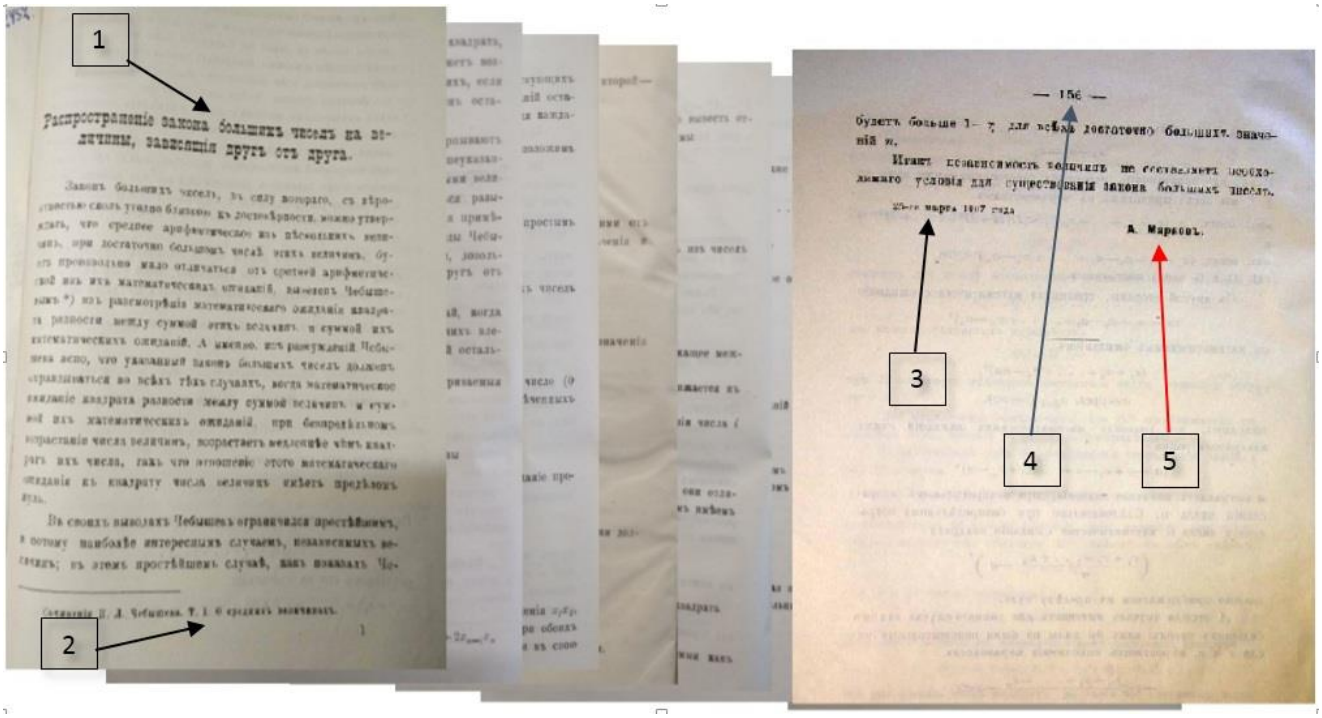


Рис. 2. Отсутствие в статьях информации, необходимой для формирования набора метаданных, в частности, сведений об авторах, ключевых слов. Структурные и стилевые особенности статей позволяют найти строки, из которых можно извлечь метаданные: 1 – название статьи, 2 – ссылка на научную статью или книгу, 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

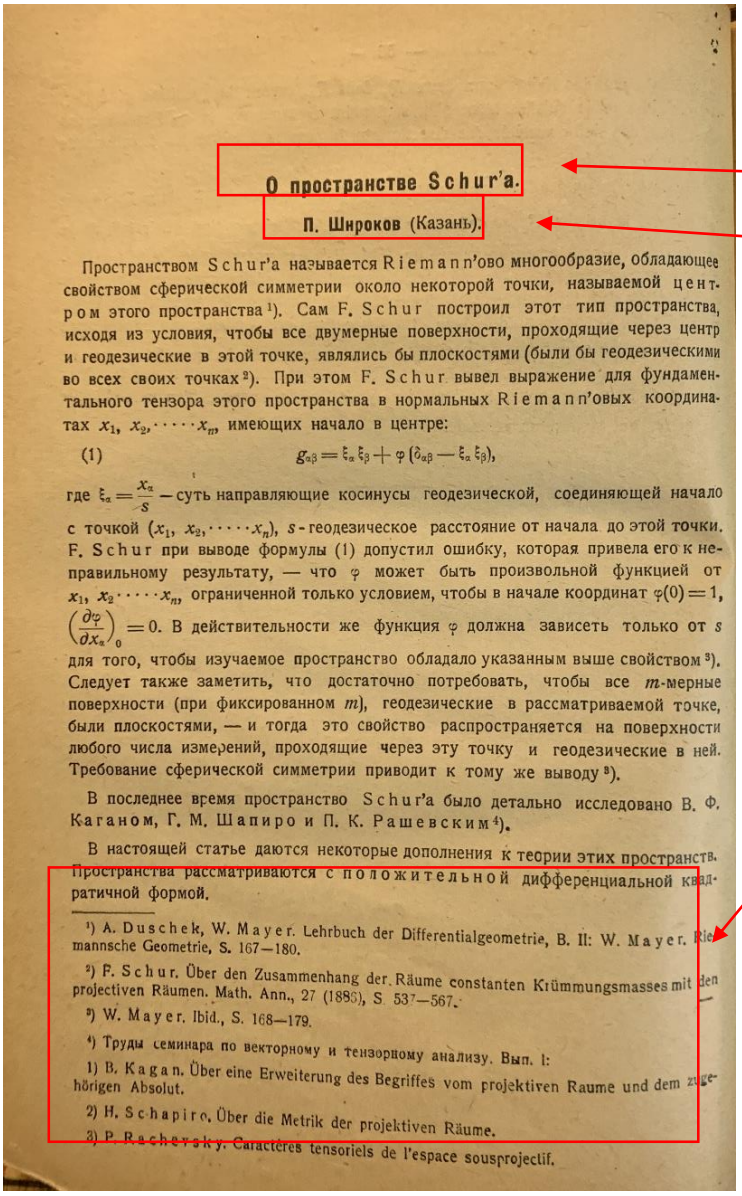
11: **осуществить поиск строки,** содержащей список авторов
12: **выделить** авторов статьи по шаблонам регулярных выражений
13: **осуществить поиск и извлечение** блока аннотации
14: **осуществить поиск и извлечение** списка литературы
15: **уточнить информацию об авторе** из открытых интернет источников
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.

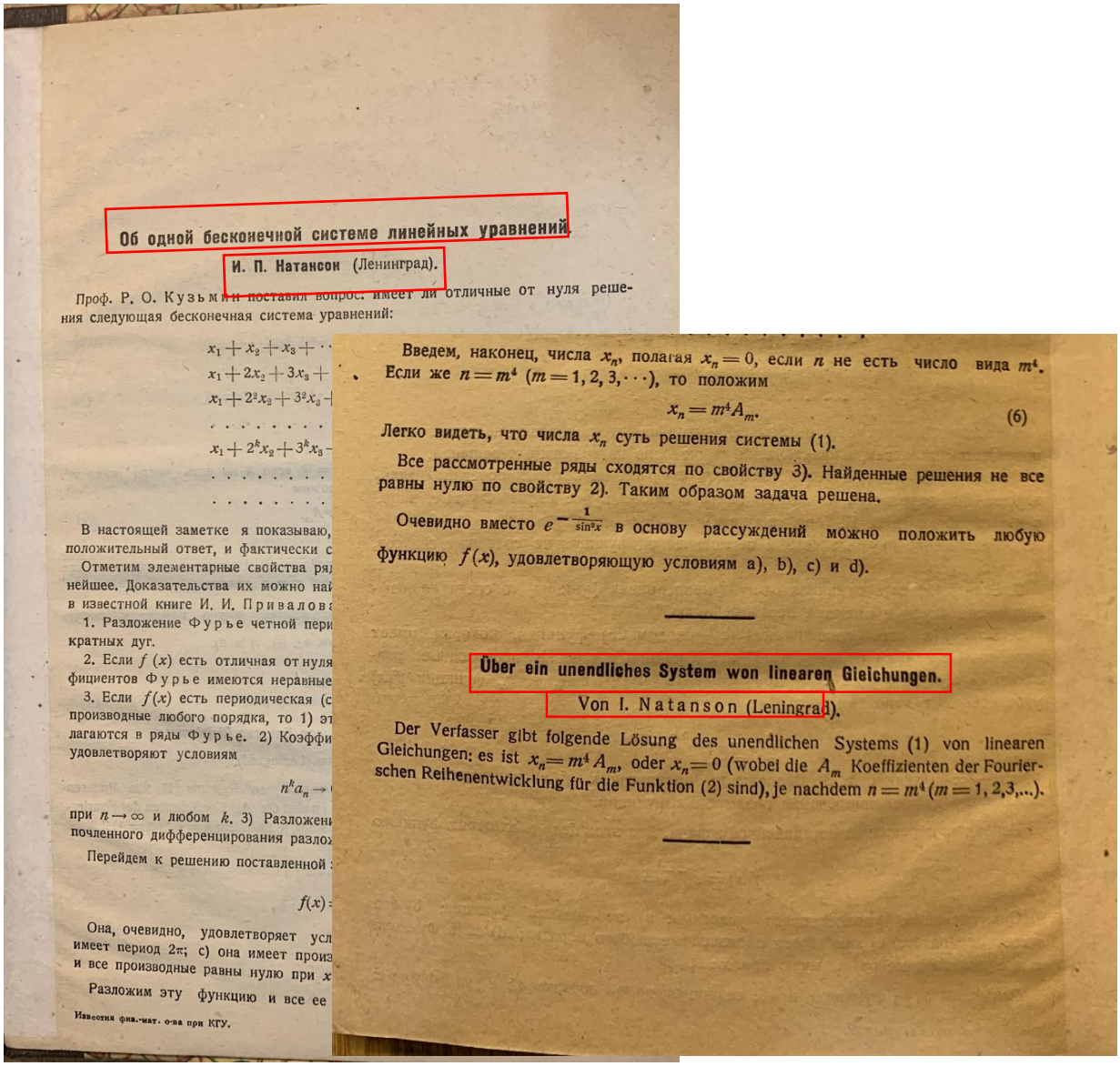
3) P. Rachevsky. Caractères tensoriels de l'espace sousprojectif.

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

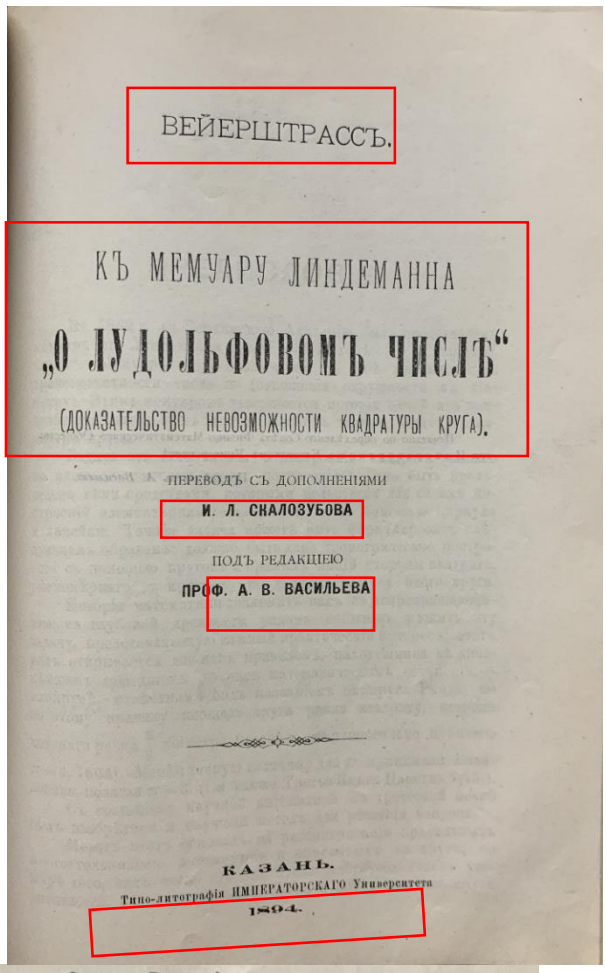
Название статьи

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

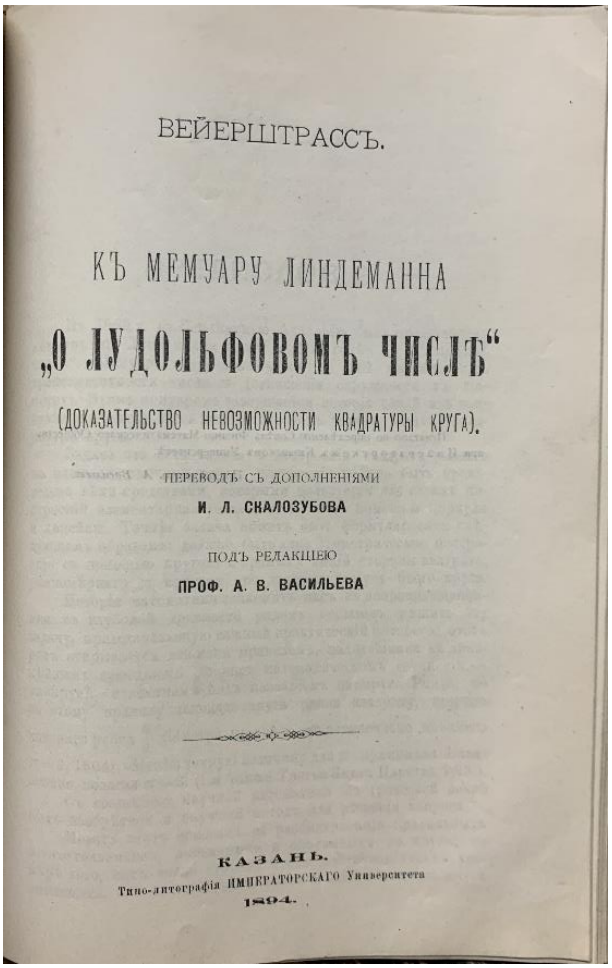
КАЗАНЬ.

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

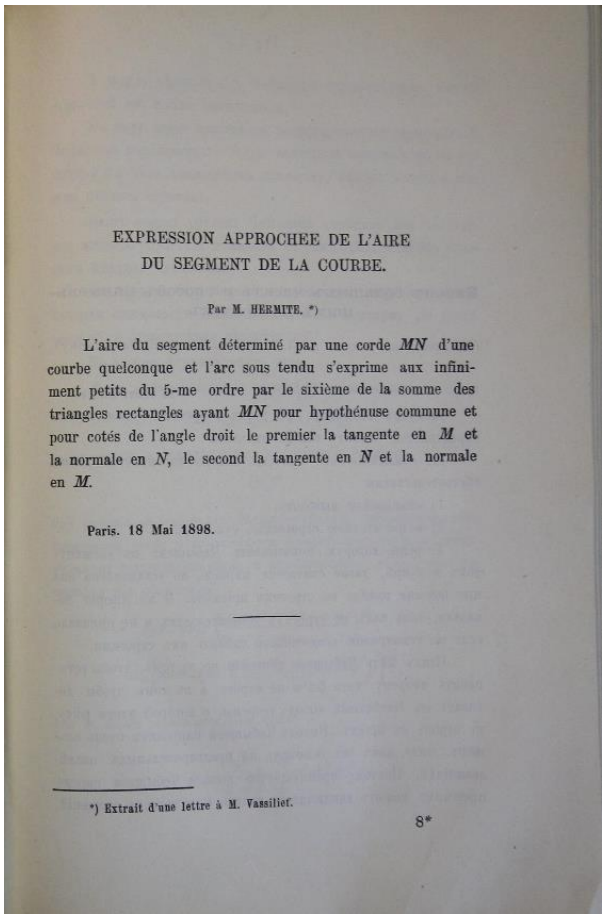
ПОДЪ РЕДАКЦІЕЮ

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.

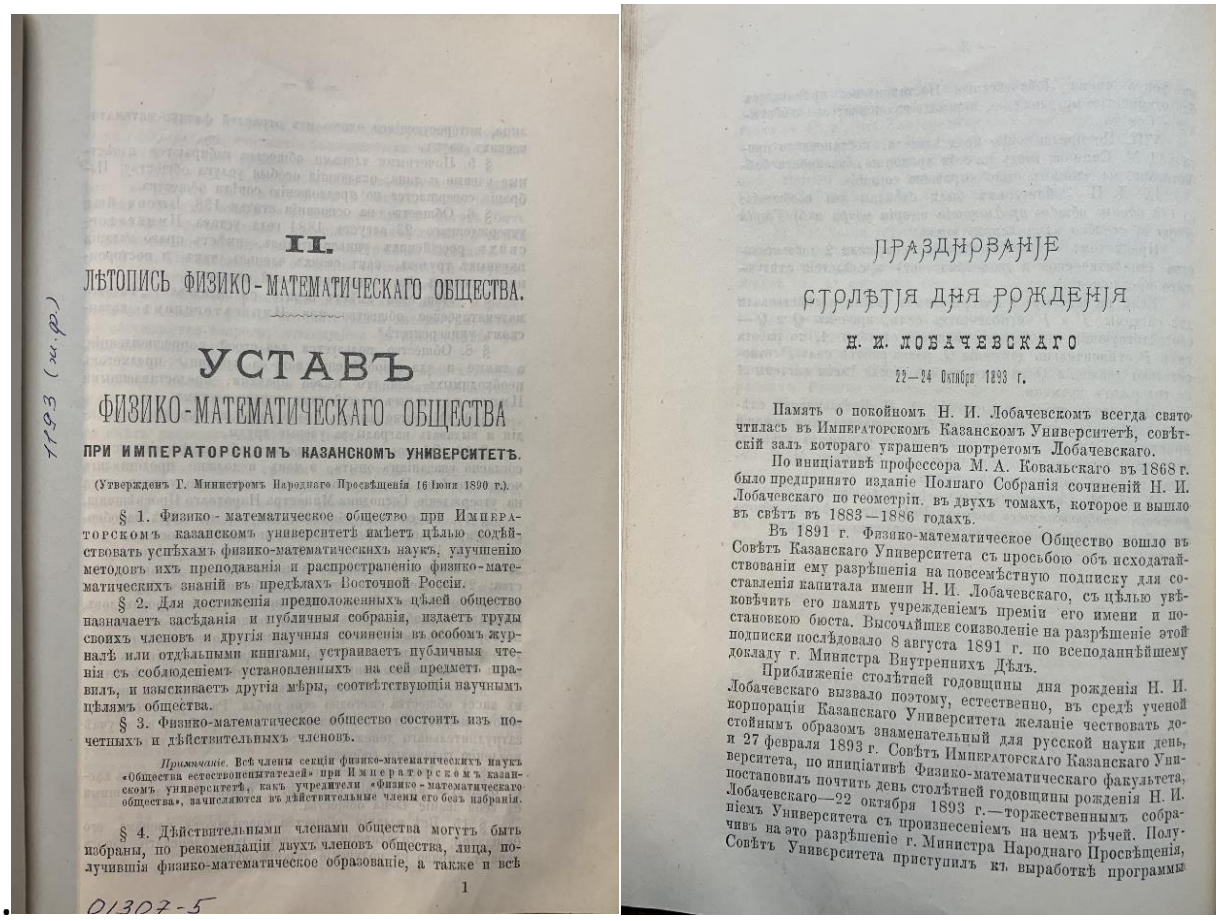
8*

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