УДК 004.4

СИСТЕМА АВТОМАТИЧЕСКОГО ФОРМИРОВАНИЯ, ОБРАБОТКИ И УПРАВЛЕНИЯ МЕТАДАННЫМИ ДОКУМЕНТОВ ЭЛЕКТРОННЫХ КОЛЛЕКЦИЙ

А. Р. Хамеджанов[0009-0000-5624-2453]

Казанский (Приволжский) федеральный университет, г. Казань, Россия

hamedzhanovalmaz@gmail.com

Аннотация

В настоящее время издательский цикл претерпевает значительные технологические изменения: внедряются автоматизированные системы управления публикационными процессами, используются нейросетевые технологии для обработки контента, активно развиваются инструменты интеллектуального анализа научных данных. Одним из ключевых трендов становится автоматизация издательского цикла, направленная на ускорение обработки рукописей, повышение качества метаописания и обеспечение совместимости информационных ресурсов. В этом контексте метаданные выступают связующим элементом для машинной обработки и навигации в пространстве научных знаний, обеспечивая структурирование информации, ее интерпретацию и интеграцию в цифровые библиотечные системы. Однако метаданные научных публикаций часто содержат ошибки, неточности или являются неполными, а их ручное формирование и уточнение требуют значительных временных затрат и не обеспечивают высокой точности. В работе представлена система автоматического формирования, обработки и управления метаданными научных документов на основе данных, полученных из сервисов поиска научных публикаций и открытых баз знаний. Эта система может использоваться для автоматизации процесса извлечения, уточнения и дополнения метаданных научных публикаций с целью последующего формирования электронных коллекций научных документов.

***Ключевые слова**: цифровая математическая библиотека, семантическая сеть, автоматизация редакционных процессов, формирование метаданных, извлечение метаданных, дополнение метаданных, NISO JATS, цифровая библиотека.*

ВВЕДЕНИЕ

Метаданные являются не просто описанием данных, но и выступают связующим звеном, обеспечивающим структурирование знаний, его интерпретацию и возможность навигации в пространстве научных знаний, включая традиционные библиотечные каталоги или современные семантические веб-системы [1]. В компьютерных системах метаданные выполняют ключевую функцию обеспечения совместимости различных информационных ресурсов и автоматизированной обработки массивов данных [2]. В контексте цифровых библиотек качество метаданных напрямую определяет эффективность поиска, доступность контента и долгосрочную сохранность научного наследия [3]. Помимо автоматизированных систем, работающих с обработкой данных, метаданные нужны также для создания пользователями запросов, анализа данных и интерпретации их содержимого. В ряде источников также отмечается ключевая роль метаданных в обеспечении технических стандартов и правил генерации записей, что значительно упрощает процесс работы с данными [4, 5].

Однако метаданные по разным причинам могут содержать ошибки и неточности (например, в случаях, когда авторы имеют одно и то же полное имя). Кроме того, ручное формирование блока метаданных требует существенных временных затрат.

В рамках проведенного исследования для решения ряда отмеченных проблем, реализована система автоматического формирования, обработки и уточнения исходных метаданных и дополнения недостающей информации на основании полученных данных с помощью поисковой системы Google Scholar (https://scholar.google.com/), систем ORCID (https://orcid.org/), Yandex Translate (https://translate.yandex.ru/) и запросов к графу знаний WikiData (https://www.wikidata.org). Разработанная система позволяет извлекать метаданные на основе научных публикаций, загруженных в нее, и формировать выходной XML-файл в формате NISO JATS V1.0 (Journal Article Tag Suite, https://jats.nlm.nih.gov/1.0). Данные, которые не были указаны в статье, например ключевые слова или элементы аффилиации, могут быть дополнены из открытых источников.

Для эффективной интеграции сервисов в функционал цифровых библиотек, а также для обеспечения их совместимости с внешними библиотечными системами и базами данных критически важно уделять внимание согласованности форматов метаданных, используемых в этих информационных ресурсах. Цифровые математические библиотеки DML-CZ (Czech Digital Mathematics Library, https://dml.cz), Numdam (http://www.numdam.org) и EuDML (The European Digital Mathematics Library, https://initiative.eudml.org) используют XML-схемы NISO JATS V1.0 для описания публикаций из математических изданий согласно международным стандартам, предложенным в проекте Всемирной цифровой математической библиотеки (World Digital Mathematical Library – WDML) [6]. Поэтому коллекции метаданных, собранные для цифровой библиотеки Lobachevskii-DML, также должны соответствовать международным стандартам, чтобы обеспечить интеграцию этих коллекций в агрегирующие научные библиотеки.

Стандарт NISO JATS версии 1.0 представляет собой набор тегов для структурирования и описания научных статей в формате XML. Этот стандарт включает в себя множество полей, которые обеспечивают детализированное описание статьи. Информацию о конкретных тегах можно найти, например, в https://jats.nlm.nih.gov/1.0. Фрагмент XML-кода, приведенный на рис. 1, иллюстрирует компоновку метаданных научной статьи в соответствии с этим стандартом.

Согласно стандартам, разработанным EuDML [7–10], наборы данных различаются по степени необходимости их включения в XML-документ. Перечислим их.

Обязательные (Mandatory) – это элементы, которые должны присутствовать в каждом JATS-документе. Они необходимы, чтобы документ соответствовал минимальным требованиям стандарта. Примеры обязательных элементов включают `<article-id>` (уникальный идентификатор статьи), `<article-title>` (название на языке оригинала) и `<contrib-group>` (список авторов).

```
<article>
    <front>
        <article-meta>
            <title-group>
                <article-title>Название статьи</article-title>
            </title-group>
            <contrib-group>
                <contrib contrib-type="author">
                    <name>
                        <surname>Хамеджанов</surname>
                        <given-names>A.P.</given-names>
                    </name>
                    <xref ref-type="aff" rid="aff0"/>
                </contrib>
                <!-- Остальные авторы -->
            </contrib-group>
            <!-- Аффиляция -->
            <aff id="aff1">
                <institution>Название университета</institution>
                <addr-line>Адрес университета</addr-line>
                <country>Страна</country>
            </aff>
            <abstract>
                <p>Аннотация статьи...</p>
            </abstract>
            <kwd-group>
                <kwd>Ключевое слово 1</kwd>
                <kwd>Ключевое слово 2</kwd>
                <!-- Дополнительные ключевые слова -->
            </kwd-group>
        </article-meta>
    </front>
</article>
```

Рис. 1. Фрагмент XML-файла, созданного согласно схеме NISO JATS V1.0.

Фундаментальные (Fundamental) – эти элементы считаются основными для структуры и содержания научной статьи, но не всегда обязательны для каждой статьи. Они включают такие элементы, как <abstract> (аннотация) и <kwd-group> (ключевые слова), которые предоставляют основную информацию о статье и ее содержании.

Дополнительные (Supplemental) – это элементы, которые могут быть добавлены в JATS-документ для обогащения информации, но не являются необходимыми для соответствия стандарту. Примеры включают `<ref-list>` (список литературы), `<funding-group>` (данные о финансировании исследования), `<ext-link>` (различные ссылки с дополнительной информацией) и др., которые могут предоставлять дополнительные данные о статье или ее авторах.

В работе [11] отмечено, что схемы, предложенные EuDML, не позволяют в рамках единого набора метаданных описать статью, опубликованную на русском языке в научном журнале, и ее перевод в англоязычной версии этого журнала. Рекомендованы расширенная xml-схема, учитывающая указанную особенность, и соответствующие алгоритмы нормализации метаданных.

Данные разграничения блоков метаданных также актуальны для текущей разрабатываемой системы. Настоящая работа является продолжением исследований, представленных в [12].

В первой части содержится краткий анализ близких по тематике научных исследований. Во второй части описаны модель работы системы, спроектированная архитектура проекта и алгоритмы извлечения, дополнения и уточнения метаданных. В последней части указаны примеры метаданных научных документов, которые можно получить при обращении к различным семантическим сетям.

1. ИССЛЕДОВАНИЯ, БЛИЗКИЕ ПО ТЕМАТИКЕ

В работе [13] предложен алгоритм автоматического формирования метаданных выпусков научного журнала для экспорта в международные информационно-аналитические системы.

В статьях [14–17] предложен метод уточнения и дополнения аффилиации авторов с использованием семантической сети WikiData. Разработаны алгоритмы извлечения аффилиации из текста и дополнения полученных метаданных через SPARQL-запросы к базе знаний WikiData. Предложенные методы позволяют автоматически уточнять информацию об организациях – местах работы авторов статей, включая названия, адреса, страны и другие атрибуты.

Как правило, ядром экосистемы цифровых библиотек является фабрика метаданных (см., например, [18, 19]). В работе [19] приведено следующее определение этого термина: фабрика метаданных – это система взаимосвязанных программных инструментов, направленных на создание, обработку, хранение и управление метаданными объектов цифровых библиотек и позволяющих интегрировать создаваемые электронные коллекции в агрегирующие цифровые научные библиотеки.

В статьях [20, 21] представлена фабрика метаданных цифровой математической библиотеки Lobachevskii-DML [22, 23]. Эта фабрика включает набор сервисов для автоматизированного формирования, обработки и верификации метаданных; реализованы также программные инструменты нормализации метаданных в форматы агрегирующих библиотек [24]. Система, представленная в настоящем исследовании, разрабатывалась в соответствии со структурой и схемами указанной фабрики метаданных.

В работах [25, 26] предложены методы автоматической обработки научных документов, основанные на анализе структуры документов и применении методов семантического анализа. В частности, разработан метод извлечения из документа именованных сущностей с использованием предметных онтологий, что позволяет расширить набор ключевых слов документа. В [27, 28] представлена система сервисов автоматической обработки больших коллекций научных документов: извлечение метаданных из документов коллекции производится на основе анализа их структуры и форматов представления информации; созданные сервисы используют онтологии описания структуры документов [29].

В [30, 31] предложены решения основных задач, связанных с формированием цифровых математических коллекций из документов, изданных в доцифровой период, – такие коллекции обозначены авторами как ретроколлекции. Разработаны алгоритмы создания метаописания ретроколлекций, основанные на анализе структуры математических документов и применении программных инструментов выделения метаданных. Представлено описание ретроколлекций, сформированных с помощью разработанных алгоритмов и включенных в состав фабрики метаданных цифровой математической библиотеки Lobachevskii-DML. Указаны схемы формирования метаданных и методы нормализации извлеченной информации в соответствии со схемами и требованиями интегрирующих математических библиотек.

2. МОДЕЛЬ РАБОТЫ ПРОГРАММНОГО РЕШЕНИЯ

Процесс работы созданной программы можно разбить на отдельные этапы, где каждый следующий этап выполняется на основе результата предыдущего. Принцип работы системы в общем виде представлен на рис. 2 в виде UML-диаграммы деятельности.

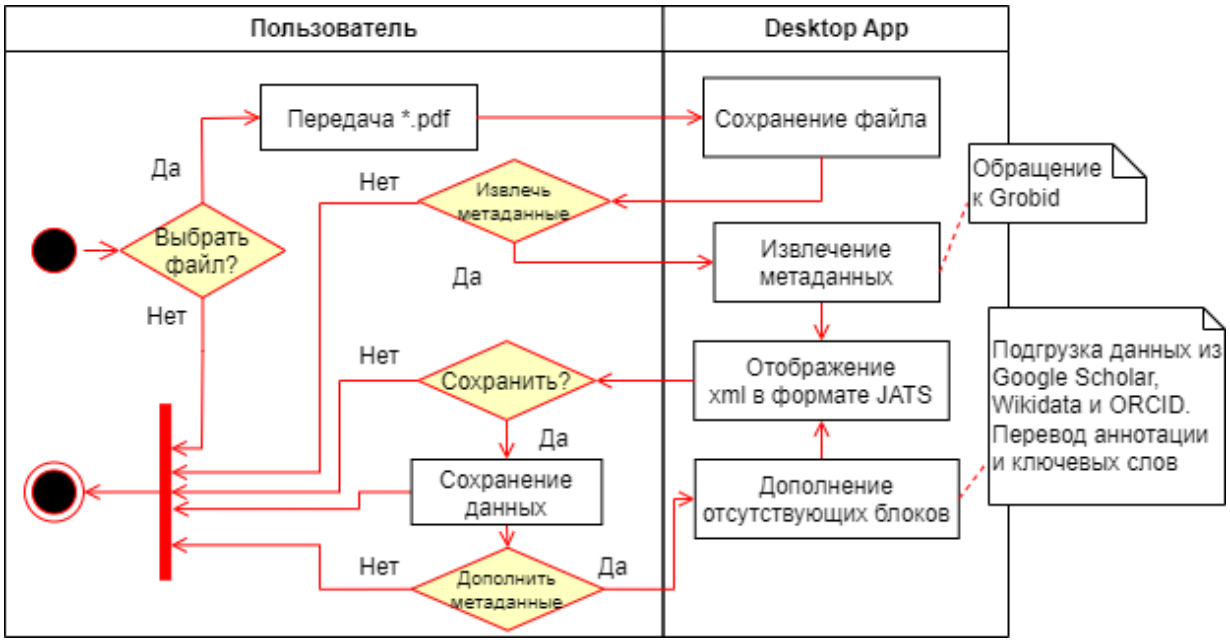

Рис. 2. Модель работы плагина в виде UML-диаграммы деятельности.

В общем виде работу системы можно описать следующим образом: пользователь подает на вход научный документ в виде pdf-файла, далее он может воспользоваться функцией извлечения метаданных, которая проанализирует документ и выделит блоки метаданных с помощью программного сервиса Grobid (GeneRation Of BIbliographic Data, https://grobid.readthedocs.io/en/latest/Introduction) и выведет данные в формате NISO JATS. На следующем шаге пользователю доступна функция дополнения метаданных, которая проанализирует ранее полученный xml-файл и выполнит поиск данных в Google Scholar, WikiData и ORCID, а также осуществит перевод аннотации и ключевых слов с помощью Yandex Translate. Полученные результаты оформляются в единый xml-файл и выводятся на экран. На каждом этапе работы пользователь может сохранить полученный промежуточный результат.

Разработанный алгоритм извлечения метаданных из научной работы включает функциональную последовательность операций (рис. 3). Сначала выбираются параметры извлечения и сохраняется файл в системе для дальнейшей передачи. Затем создается и отправляется запрос к сервису извлечения метаданных для обработки файла согласно заданным настройкам с помощью специального http-клиента Ktor (https://ktor.io). Далее данные, полученные в формате TEI (Text Encoding Initiative), преобразуются в соответствии с XML-схемой NISO JATS и специальным XSL-файлом (eXtensible Stylesheet Language) в XML-файл.

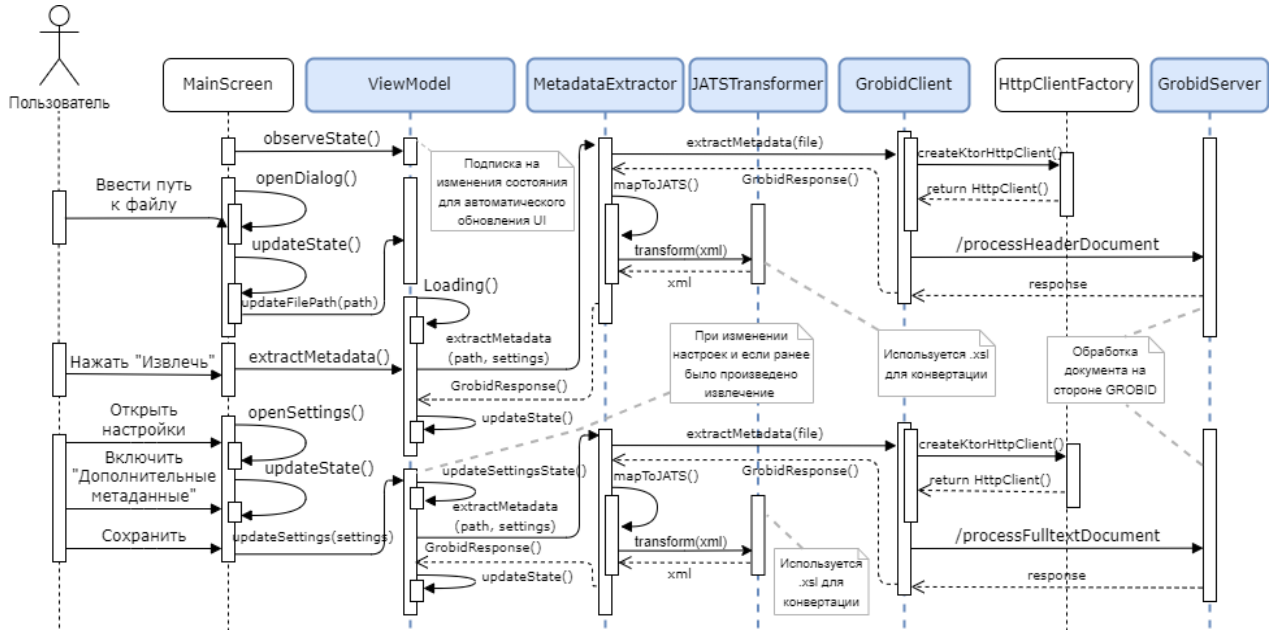

Рис. 3. Диаграмма последовательности извлечения метаданных из научного документа с настройкой необходимости формирования и включения в набор дополнительных метаданных.

Общая схема процесса дополнения и уточнения блоков метаданных состоит из нескольких шагов и представлена в виде псевдокода ниже (см. Алгоритм 1). На первом этапе выполняется поиск метаданных по названию статьи и списку авторов в поисковой системе Google Scholar. Далее проводится обогащение полученной информации посредством обращения к базе знаний WikiData. На следующем шаге осуществляется поиск по названию статьи в реестре уникальных идентификаторов авторов ORCID. В случае обнаружения кода ORCID автора при извлечении метаданных из документа либо на предыдущих шагах выполняется обращение к реестру ORCID для уточнения и дополнения информации

об авторе. Завершающим этапом являются анализ и сбор блоков метаданных в единый XML-файл согласно схеме NISO JATS.

Алгоритм 1: Дополнение и уточнение блоков метаданных из открытых источников

```
# Получение данных по извлеченному названию научной статьи
article_data = google_scholar.search_by_title(article_title)
# Формирование единого множества авторов
authors_set = parsed_authors.merge()
# Поиск данных по каждому автору согласно списку ФИО и id профилей авторов
author_profiles = []
for author in authors_set:
   if author.profile_id != null:
      author_details = google_scholar.get_author_profile(author.profile_id)
      author_profiles.add(author_details)
   else:
      author_profiles.add(google_scholar.search_author(author.full_name))
end for
# Получение данных для цитирования по уникальному идентификатору статьи
citation_data = google_scholar.get_citation_data(article_data.article_id)
# Объединение метаданных в единую модель
scholar_data = merge(article_data, author_profiles, citation_data)
scholar_data = filter_by_settings(scholar_data, user_settings)

# Обогащение полученной информации через WikiData
wikidata_data = []
for author in authors_set:
   wikidata_data.add(wikidata.query(author))
end for
wikidata_data = filter_by_settings(wikidata_data, user_settings)

# Поиск по названию статьи в реестре ORCID
orcid_search_results = orcid.search_by_title(article_title)
# Формирование единого множества найденных ORCID авторов
orcid_set = find_orcid(authors_set, orcid_search_results)
orcid_authors = []
for id in orcid_set:
```

```
    # Обращение к реестру ORCID для уточнения и дополнения данных об авторе
    orcid_authors.add(orcid.get_record(id))
end for
# Фильтрация согласно настройкам пользователя (включение/исключение дополнительных метаданных)
orcid_authors = filter_by_settings(orcid_authors, user_settings)

# Анализ и сбор блоков метаданных в единый XML-файл согласно схеме NISO JATS
result = merge(scholar_data, wikidata_data, orcid_search_results, orcid_authors)
jats_xml = format_result(result)
```

Для перевода аннотации и ключевых слов статьи с русского языка на английский (или наоборот) реализован модуль перевода с помощью сервиса Yandex Translate API. Разработанный алгоритм перевода представлен ниже в виде псевдокода (см. Алгоритм 2). Сначала выполняется запрос, содержащий название статьи, к Yandex Translate API для определения языка текста. Далее отправляются два запроса с текстом аннотации и списком ключевых слов. Полученные результаты перевода интегрируются в выходной XML-файл согласно стандарту NISO JATS V1.0.

Алгоритм 2. Перевод аннотации и ключевых слов статьи

```
# Разбор XML-дерева и получение названия статьи
load article_title = input.xml
# Разбор XML-дерева и получение текста аннотации
load abstract = input.xml
# Разбор XML-дерева и получение списка ключевых слов
load kwds = input.xml
# Определение языка по названию статьи
article_lang = yandex_api.detect(article_title)
# Перевод на английский, если текст написан на русском
if article_lang == 'ru':
    # Конфигурация и отправка запроса перевода на английский
    if abstract != "":
        abstract_translated = translate_en(abstract)
    for kwd in kwds:
```

```
    if kwd != "":
      kwds_translated = translate_en(kwds)
  end for
# Перевод на русский, если текст написан на английском
if article_lang == 'en':
  # Конфигурация и отправка запроса перевода на русский
  if abstract != "":
    abstract_translated = translate_ru(abstract)
  for kwd in kwds:
    if kwd != "":
      kwds_translated = translate_ru(kwds)
  end for
# Интеграция результатов с языковой меткой в xml файл

add_translation(article_lang, abstract_translated, kwds_translated)
```

Архитектура разработанной системы представлена в виде диаграммы компонентов, которая описывает связи внутри программного решения (рис. 4). Такое архитектурное решение убирает связанность между модулями, что, в свою очередь, позволяет в дальнейшем интегрировать дополнительные источники данных для увеличения полноты и точности формируемых метаданных.

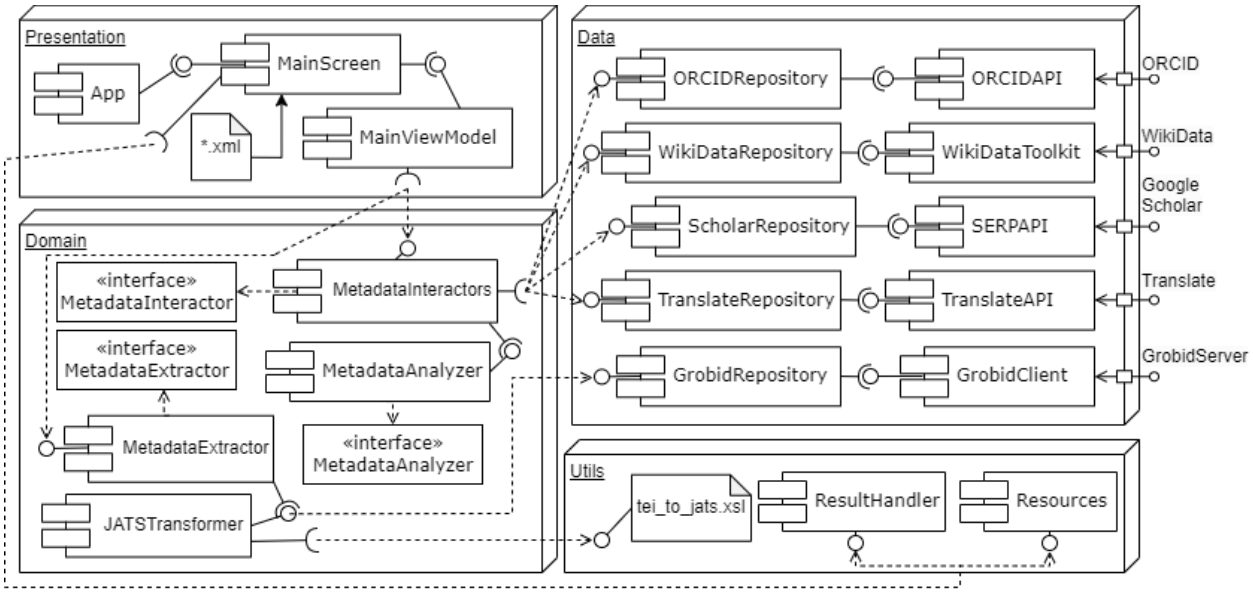

Рис. 4. Диаграмма компонентов разработанной системы автоматического формирования блока метаданных научных документов с использованием открытых баз данных.

3. РЕЗУЛЬТАТЫ

На рис. 5–7 представлен пример фрагмента ответа от сервиса Google Scholar SERP API в формате JSON (ссылки и аннотации были сокращены для читабельности примеров). Согласно этим данным (рис. 5) можно получить аффилиации, фотографию из профиля автора и список ключевых тематик работ автора, который в дальнейшем может помочь решить проблему отличия полных тезок. Кроме того можно расширить блок дополнительных метаданных ссылкой на публикацию, названием журнала и издательской компании, годом публикации, аннотацией и отдельной ссылкой на полный текст статьи (рис. 6 и 7).


```
"profiles": [
  {
    "name": "Evgeny Lipachev",
    "author_id": "HWLef7EAAAAJ",
    "affiliations": "Kazan Federal University",
    "email": " elipachev@gmail.com",
    "cited_by": 1211,
    "interests": [
      {
        "title": "Веб-технологии"
      },
      {
        "title": "краевые задачи дифракции"
      },
      {
        "title": "электронные библиотеки"
      },
      {
        "title": "MathML"
      }
    ],
"thumbnail":"https://scholar.googleusercontent.com/citations?view
_op=small_photo&user=HWLef7EAAAAJ&citpid=2"
```


Рис. 5. Фрагмент ответа от сервиса Google Scholar SERP API при поиске по ФИО автора "Evgeny Lipachev".

```
      "title": "OntoMath PRO Ontology: A Linked Data Hub for Mathematics",
      "result_id": "BcazuB-eH48J",
      "link": "https://link.springer.com/chapter/10.1007/978-3",
      "snippet": "In this paper, we present an ontology of...",
      "publication_info": {
        "summary": "OA Nevzorova, N Zhiltsov, A Kirillovich… - … Engineering and
the …, 2014 - Springer",
       "authors": [{
           "name": "OA Nevzorova",
           "link":"/citations?user=n2GFYqkAAAAJr&hl=en&oi=sra,
           "author_id": "n2GFYqkAAAAJ"
         },//остальные авторы ]
      },
      "resources": [//ссылки для скачивания]
```


Рис. 6. Фрагмент ответа от сервиса Google Scholar SERP API при поиске по названию статьи "OntoMath PRO Ontology: A Linked Data Hub for Mathematics".


```
          "citations": [
              {
                "title": "MLA",
                   "snippet": "Nevzorova, Olga A., et al. \"OntoMath PRO
ontology: a linked data hub for mathematics.\" Knowledge Engineering
and the Semantic Web: 5th International Conference, KESW 2014, Kazan,
Russia, September 29-October 1, 2014. Proceedings 5. Springer
International Publishing, 2014."
              },
```


Рис. 7. Пример фрагмента ответа от сервиса Google Scholar SERP API запросе информации для цитирования статьи по ранее найденному уникальному идентификатору "BcazuB-eH48J" (см. рис. 5).

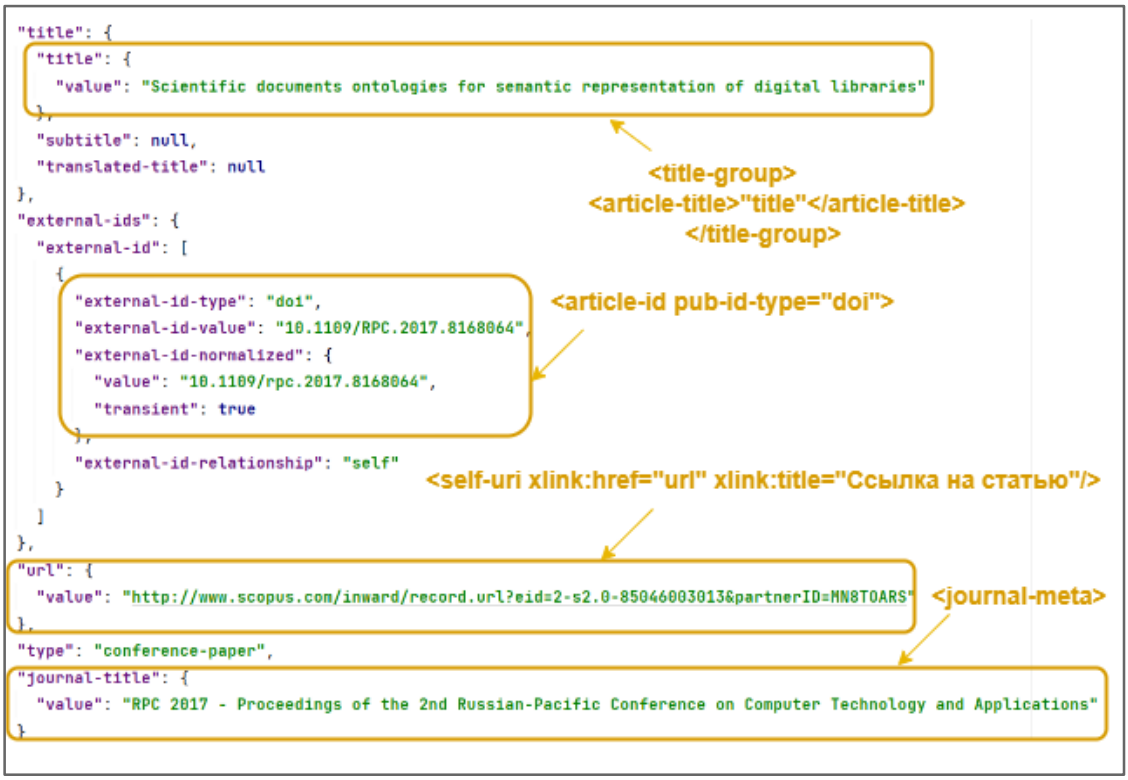

Рис. 8. Фрагмент ответа от сервиса ORCID API при запросе данных по коду автора «0000-0001-7789-2332».

Извлечение информации об авторе из реестра ORCID позволяет получить данные не только об авторе, но и о текущей статье, если в ней они присутствуют. Для этого выполняется поиск статьи среди списка публикаций автора, полученного на предыдущей шаге. Подобный подход позволяет установить блок дополнительных метаданных при их наличии, таких как DOI статьи, перевод заголовка статьи, ссылку на публикацию в журнале, дату публикации и название журнала (рис. 8).

ЗАКЛЮЧЕНИЕ

Благодаря взаимодействию с Google Scholar, ORCID и WikiData могут быть уточнены и дополнены аффилиации, ФИО, адрес электронной почты, код ORCID автора, а также дополнительные метаданные в виде ссылок на профили авторов в других научных сервисах, места работы и учебы авторов, списки научных статей, соответствующих ключевых слов, тематика научных работ. Многое зависит от степени открытости профиля авторов (https://info.orcid.org/privacy-policy),

а также от полноты информации, представленной на сайте ORCID. Работа с данным сервисом является полезным инструментом, так как в некоторых случаях подобная информация позволит не просто уточнить и дополнить метаданные, но и расширить круг поиска.

Основным результатом работы является разработка гибкой и расширяемой архитектуры системы, где каждый модуль инкапсулирует строго определенную задачу и может быть независимо заменен или модернизирован без нарушения целостности остальных компонентов. Дальнейшее направление развития заключается в совершенствовании отдельных модулей, отвечающих за методы обработки, извлечения метаданных и добавлении новых источников информации.

СПИСОК ЛИТЕРАТУРЫ

1. *Gartner R.* Metadata. Shaping Knowledge from Antiquity to the Semantic Web. Springer Cham, 2016. https://doi.org/10.1007/978-3-319-40893-4

2. *Kogalovsky M.R.* Metadata in Computer Systems // Programming and Computer Software. 2013. V. 39, No 4. P. 182–193. https://doi.org/10.1134/S0361768813040038

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

4. *Когаловский М.Р.* Метаданные, их свойства, функции, классификация и средства представления // CEUR Workshop Proceedings. 2012. V. 934. P. 3–14.

5. *Когаловский М.Р., Серебряков В.А.* Метаданные // Общенациональный интерактивный энциклопедический портал «Знания». 2022. № 9. https://doi.org/10.54972/00000048_2022_9_48

6. *Olver P.J.* The World Digital Mathematics Library: Report of a Panel Discussion // Proceedings of the International Congress of Mathematicians, August 13–21, 2014, Seoul, Korea. Kyung Moon SA, 1. 2014. P. 773–785.

7. EuDML metadata schema specification (v2.0–final). URL: https://initiative.eudml.org/eudml-metadata-schema-specification-v20-final (дата доступа 14.03.2026)

8. The EuDML metadata schema. Revision: 1.6 as of 15th December 2010.

/ Jost M., Bouche T., Goutorbe C., Jorda J.P. URL: http://www.mathdoc.fr/publis/d3.2-v1.6.pdf (дата доступа 14.03.2026)

9. *Sylwestrzak W., Borbinha J., Bouche T., Nowiński A., Sojka P.* EuDML – Towards the European Digital Mathematics Library // In: SojkaP. (Ed.) Towards a Digital Mathematics Library. Masaryk University, 2010. P. 11–26.
URL: https://eudml.org/doc/220786 (дата доступа 14.03.2026)

10. *Bouche T.* Reviving the free public scientific library in the digital age? The EuDML project // In: Kaiser K., Krantz S.G., Wegner B. (Eds.) Topics and Issues in Electronic Publishing JMM/AMS Special Session, FIZ Karlsruhe. 2013. P. 57–80.
URL: https://www.emis.de/proceedings/TIEP2013/05bouche.pdf (дата доступа 14.03.2026)

11. *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.

12. *Хамеджанов А.Р.* Система автоматического формирования блока метаданных научных документов с использованием открытых баз данных // Системы высокой доступности. 2026. Т. 22, № 1. С. 51−55.
https://doi.org/ 10.18127/j20729472-202601-10

13. *Герасимов А.Н., Елизаров А.М., Липачев Е.К.* Формирование метаданных для международных баз цитирования в системе управления электронными научными журналами // Электронные библиотеки. 2015. Т. 18, № 1–2. С. 6–31.

14. *Гафурова П.О., Липачев Е.К.* Метод уточнения аффилиации авторов научных документов на основе запросов к семантической сети // Научный сервис в сети Интернет: труды XXIV Всероссийской научной конференции. М.: ИПМ им. М.В. Келдыша, 2022. С. 115–127. https://doi.org/10.20948/abrau-2022-31

15. *Gafurova P., Elizarov A., Lipachev E.* Algorithms for Integration of Unstructured Mathematical Documents into the Common Digital Space of Scientific Knowledge // Proc. Int. Conf. «Common Digital Space of Scientific Knowledge: Problems & Solutions« (CDSSK–2020). Moscow, Russia, November 10–12, 2020. CEUR Workshop Proceedings. 2021. V. 2990. P. 39–49.
http://ceur-ws.org/Vol-2990/rpaper4.pdf

16. *Elizarov A., Gafurova P., Lipachev E.* Wikidata in Metadata Formation

Methods for Documents of Digital Mathematical Library // CEUR Workshop Proceedings. 2021. V. 3066. P. 23–33.

17. *Гафурова П.О., Елизаров А.М., Липачев Е.К.* Извлечение знаний из Wikidata для формирования метаданных документов электронных математических коллекций // Электронные библиотеки. 2021. Т. 24, № 6. С. 1023–1059. https://doi.org/10.26907/1562-5419-2021-24-6-1023-1059

18. *Bouche T., Labbe O.* The New Numdam Platform // In: Geuvers H., England M., Hasan O., Rabe F., Teschke O. (Eds.) Intelligent Computer Mathematics. CICM 2017. Lecture Notes in Computer Science. Vol. 10383. Springer, Cham, 2017. P. 70–82. https://doi.org/10.1007/978-3-319-62075-6_6

19. *Елизаров А.М., Липачев Е.К.* Цифровые платформы и цифровые научные библиотеки // International Journal of Open Information Technologies. 2020. Т. 8. № 11. С. 80–90.

20. *Elizarov A., Lipachev E.* Digital Library Metadata Factories // CEUR Workshop Proceedings. 2021. V. 2813. P. 13–21.

21. *Гафурова П.О., Елизаров А.М., Липачев 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

22. *Елизаров А.М., Липачев Е.К.* Цифровая библиотека Lobachevskii-DML в научном пространстве математических знаний // Научно-техническая информация. Серия 1: Организация и методика информационной работы. 2023. № 1. С. 32–37. https://doi.org/10.36535/0548-0019-2023-01-3

23. *Elizarov A., Lipachev E.* BIG MATH Methods in Lobachevskii-DML Digital Library // CEUR Workshop Proceedings. 2019. V. 2523. P. 59–72.

24. *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

25. *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. https://doi.org/10.3103/S000510551402006X

26. *Биряльцев Е.В., Елизаров А.М., Жильцов Н.Г., Липачев 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.

27. *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. V. 1752. P. 58–68.

28. *Elizarov A., Khaydarov S., Lipachev E.* Scientific Documents Ontologies for Semantic Representation of Digital Libraries // RPC 2017 – Proceedings of the 2nd Russian-Pacific Conference on Computer Technology and Applications. 2017. P. 1–5. https://doi.org/10.1109/RPC.2017.8168064

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

30. *Андреичев М.Д., Гафурова П.О., Елизаров А.М., Липачев 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

31. *Гафурова П.О., Елизаров А.М., Липачев 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

THE SYSTEM FOR THE AUTOMATIC GENERATION, PROCESSING, AND MANAGEMENT OF DOCUMENT METADATA IN DIGITAL COLLECTIONS

A. R. Khamedzhanov [0009-0000-5624-2453]

Kazan (Volga region) Federal University, Kazan, Russia

hamedzhanovalmaz@gmail.com

Abstract

The publishing cycle is currently undergoing significant technological changes: automated publication management systems are being implemented, neural network technologies are being used for content processing, and tools for the intelligent analysis of scientific data are being actively developed. One of the key trends is the automation of the publishing cycle, aimed at accelerating manuscript processing, improving the quality of metadata, and ensuring the interoperability of information resources. In this context, metadata serves as a connecting element for machine processing and navigation within the scientific knowledge space, ensuring the structuring, interpretation, and integration of information into digital library systems. However, metadata for scientific publications often contain errors, inaccuracies, or are incomplete, and their manual creation and refinement are time-consuming and do not ensure high accuracy. The aim of this work is to design and develop a system for the automatic generation, processing, and management of metadata for scientific documents based on data obtained from scientific publication search services and open knowledge bases. The system can be used to automate the process of extracting, refining, and supplementing the metadata of scientific publications for the purpose of subsequently creating electronic collections of scientific documents.

Keywords*: digital mathematical library, semantic networks, automation of editorial processes, metadata generation, metadata extraction, metadata addition, NISO JATS, digital libraries.*

REFERENCES

1. *Gartner R.* Metadata. Shaping Knowledge from Antiquity to the Semantic Web. Springer Cham, 2016. https://doi.org/10.1007/978-3-319-40893-4

2. *Kogalovsky M.R.* Metadata in Computer Systems // Programming and Computer Software. 2013. V. 39, No. 4. P. 182–193.
https://doi.org/10.1134/S0361768813040038

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

4. *Kogalovsky M.R.* Metadata, their Properties, Functions and Classifications // CEUR Workshop Proceedings. 2012. V. 934. P. 3–14.

5. *Kogalovsky M.R., Serebryakov V.A.* Metadata // National Interactive Encyclopedia Portal "Knowledge". 2022. No. 9.
https://doi.org/10.54972/00000048_2022_9_48

6. *Olver P.J.* The World Digital Mathematics Library: Report of a Panel Discussion // Proceedings of the International Congress of Mathematicians, August 13–21, 2014, Seoul, Korea. Kyung Moon SA, 1. 2014. P. 773–785.

7. EuDML metadata schema specification (v2.0–final). URL: https://initiative.eudml.org/eudml-metadata-schema-specification-v20-final.

8. The EuDML metadata schema. Revision: 1.6 as of 15th December 2010. / Jost M., Bouche T., Goutorbe C., Jorda J.P.
URL: http://www.mathdoc.fr/ publis/d3.2-v1.6.pdf (last access 04.04.2026)

9. *Sylwestrzak W., Borbinha J., Bouche T., Nowiński A., Sojka P.* EuDML – Towards the European Digital Mathematics Library // In: SojkaP. (Ed.) Towards a Digital Mathematics Library. Masaryk University, 2010. P. 11–26.
URL: https://eudml.org/doc/220786 (last access 04.04.2026)

10. *Bouche T.* Reviving the free public scientific library in the digital age? The EuDML project // In: Kaiser K., Krantz S.G., Wegner B. (Eds.) Topics and Issues in Electronic Publishing JMM/AMS Special Session, FIZ Karlsruhe. 2013. P. 57–80.
URL: https://www.emis.de/proceedings/TIEP2013/05bouche.pdf (last access 04.04.2026)

11. *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.

12. *Khamedzhanov A.R.* The system of automatic generation of a block of metadata of scientific documents using open databases // Highly Available Systems. 2026. V. 22, No. 1. P. 51−55. https://doi.org/10.18127/j20729472-202601-10

13. *Gerasimov A.N., Elizarov A.M., Lipachev E.K*. Formation of metadata for international citation databases in the management system of electronic scientific journals // Russian Digital Libraries Journal. 2015. V. 18, No. 1–2. P. 6–31.

14. *Gafurova P.O., Lipachov E.K.* Method for Clarifying the Affiliation of Authors of Scientific Documents Based on Requests to the Semantic Web. XXIV All-Russian Scientific Conference 'Scientific Service on the Internet'. 2022. P. 115–127. https://doi.org/10.20948/abrau-2022-31

15. *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. V. 2990. P. 39–49.
URL: http://ceur-ws.org/Vol-2990/rpaper4.pdf (last access 04.04.2026)

16. *Elizarov A., Gafurova P., Lipachev E.,* Wikidata in Metadata Formation Methods for Documents of Digital Mathematical Library // CEUR Workshop Proceedings. 2021. V. 3066. P. 23–33.

17. *Gafurova P.O., Elizarov A.M., Lipachev E.K.* Extraction of Wikidata Knowledge for the Metadata Formation for Documents of Electronic Mathematical Collections // Russian Digital Libraries Journal. 2021. V. 24, No. 6. P. 1023–1059.
https://doi.org/10.26907/1562-5419-2021-24-6-1023-1059

18. *Bouche T., Labbe O.* The New Numdam Platform // In: Geuvers H., England M., Hasan O., Rabe F., Teschke O. (Eds.) Intelligent Computer Mathematics. CICM 2017. Lecture Notes in Computer Science. Vol. 10383. Springer, Cham, 2017. P. 70–82. https://doi.org/10.1007/978-3-319-62075-6_6

19. *Elizarov A., Lipachev E.* Digital Platforms and Digital Scientific Libraries // International Journal of Open Information Technologies. 2020. V. 8, No. 11. P. 80–90.

20. *Elizarov A., Lipachev E.* Digital Library Metadata Factories // CEUR Workshop Proceedings. 2021. V. 2813. P. 13-21.

21. *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

22. *Elizarov A.M., Lipachev E.K.* Lobachevskii Digital Library in the Scientific Space of Mathematical Knowledge // Automatic Documentation and Mathematical

Linguistics Series 1: Organization and Methods of Information Work. 2023. No. 1. P. 32–37. https://doi.org/10.36535/0548-0019-2023-01-3

23. *Elizarov A., Lipachev E.* BIG MATH Methods in Lobachevskii-DML Digital Library // CEUR Workshop Proceedings. 2019. V. 2523. P. 59–72.

24. *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

25. *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. https://doi.org/10.3103/S000510551402006X

26. *Biryal'tsev E., Elizarov A., Zhil'tsov N., Lipachev E., Nevzorova O., Solov'ev V.* Methods for analyzing semantic data of mathematical electronic collections // Scientific and Technical Information. Series 2: Information Processes and Systems. 2014. No 4. P. 12–17.

27. *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. V. 1752. P. 58–68.

28. *Elizarov A., Khaydarov S., Lipachev E.* Scientific Documents Ontologies for Semantic Representation of Digital Libraries // RPC 2017 – Proceedings of the 2nd Russian-Pacific Conference on Computer Technology and Applications. 2017. P. 1–5. https://doi.org/10.1109/RPC.2017.8168064

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

30. *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. XXIII All-Russian Scientific Conference 'Scientific Service on the Internet'. 2021. P. 22–33. https://doi.org/10.20948/abrau-2021-22

31. *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. V. 24, No 2. P. 238–271. https://doi.org/10.26907/1562-5419-2021-24-2-238-270

СВЕДЕНИЯ ОБ АВТОРЕ

ХАМЕДЖАНОВ Алмаз Рустамович – аспирант Института информационных технологий и интеллектуальных систем Казанского федерального университета

Almaz Rustamovich KHAMEDZHANOV– postgraduate student at the Institute Institute of Information Technology and Intelligent Systems, Kazan Federal University

email: hamedzhanovalmaz@gmail.com

ORCID: 0009-0000-5624-2453

Материал поступил в редакцию 23 марта 2026 года