УДК 004.4

МЕТОД АВТОМАТИЧЕСКОГО ПОПОЛНЕНИЯ МЕТАДАННЫХ ЭЛЕКТРОННЫХ КОЛЛЕКЦИЙ ЦИФРОВОЙ МАТЕМАТИЧЕСКОЙ БИБЛИОТЕКИ

П. О. Гафурова[0000-0002-1544-155X]

Национальный исследовательский центр «Курчатовский институт», Лаборатория суперкомпьютерного моделирования, ул. Оренбургский тракт, 20А, г. Казань, 420059

pogafurova@gmail.com

Аннотация

Предложены подходы к дополнению метаданных документов электронных коллекций цифровой математической библиотеки. В качестве источников информации для пополнения метаданных использованы открытые ресурсы семантической сети. Для этой цели разработаны программные инструменты, обеспечивающие поиск необходимых данных и их включение в набор метаданных.

Предложен алгоритм пополнения метаданных аффилиации авторов научной статьи. Как правило, представленная в документе аффилиация содержит недостаточное количество информации, необходимой для формирования набора метаданных. Метод пополнения метаданных аффилиации авторов основан на данных, извлеченных из открытого реестра идентификаторов научных организаций Research Organization Registry (ROR). Также в методе использованы разработанные инструменты извлечения связей между ROR и открытыми семантическими сетями. Этот метод апробирован на электронной коллекции статей журнала «Электронные библиотеки» за 2021–2022 годы.

На основе предложенного метода разработан программный сервис, включенный в фабрику метаданных цифровой библиотеки Lobachevskii-DML. Также результатом работы является включение в цифровую библиотеку Lobachevskii-DML новых электронных коллекций. Кроме того, описан метод трансформации метаданных в формат, доступный для загрузки в библиотеку.

***Ключевые слова**:* ROR, Wikidata, цифровые библиотеки, метаданные аффилиации, Lobachevskii-DML.

ВВЕДЕНИЕ

Разработка цифровой математической библиотеки сопровождается решением задач по формированию метаданных электронных коллекций научных документов. Одной из задач, решаемых в цифровой математической библиотеке Lobachevskii-DML (https://lobachevskii-dml.ru/), является разработка методов формирования метаданных документов электронных коллекций [1]. В Lobachevskii-DML разрабатываются системы сервисов, соответствующие концепции BigMath [2, 3]. В концепции BigMath предложена модель Tetrapod [4, 5]. Модель Tetrapod включает в себя такие основные аспекты математики, как: вывод (inference), вычисление (computation), табуляция (tabulation), описание (narration), организация (organization). В библиотеке Lobachevskii-DML ставятся задачи пополнения метаданных электронных коллекций и гармонизации коллекций метаданных, которые соответствуют аспекту «Tabulation» [6–8]. Данный аспект подразумевает создание, сбор, поддержание и доступ к наборам объектов, закономерности и отношения между объектами и позволяет проверять гипотезы [4].

При включении коллекций в информационное пространство, в частности, в агрегирующие библиотеки, необходимо учитывать требования к составу метаданных [9–11]. По правилам, принятым в агрегирующих библиотеках и научных журналах, метаданные документов цифровых коллекций должны включать метаданные аффилиации. Аффилиация автора – это метаданные о месте работы автора документа [12, 13]. Метаданные аффилиации автора, как правило, требуют пополнения и уточнения. Например, часто используются сокращения названий научных организаций, может быть указана неполная информация об организации или организация не указана совсем. Разбор примеров неполных данных приведен в разделе «Особенности обработки аффилиации научных документов».

При формировании метаданных электронных коллекций цифровой библиотеки Lobachevskii-DML задача автоматического извлечения и пополнения составляющих аффилиации авторов публикаций является одной из наиболее сложных [14, 15]. В частности, необходимо точно определять личность автора статьи, что не всегда возможно. В существенной части документов электронных коллекций

присутствуют минимальные сведения об авторе и организации, что не позволяет без дополнительных сервисов и ручной обработки составить полную аффилиацию авторов документов. Решение задачи извлечения блока аффилиации сопровождается реализацией методов выделения структуры документов (см. [17]).

Пополнение метаданных цифровых коллекций можно осуществлять различными способами. Одним из таких способов является использование семантических сетей в интернете (примеры использования таких сетей приведены в [7, 16, 17]). Одними из главных особенностей семантических сетей являются упорядоченная структура сущностей и наличие точки доступа к семантической сети. Примерами таких семантический сетей являются Wikidata (https://www.wikidata.org/) и DBpedia (https://www.dbpedia.org/) [18]. Алгоритмы и методы формирования запросов к семантической сети приведены в [6, 17]. Другим подходом к дополнению метаданных является поиск в специализируемых семантических ресурсах. В данной статье для этой цели использован открытый реестр идентификаторов ROR (The Research Organization Registry) (https://ror.org/) [18, 19]. Использование специализированных коллекций улучшает качество дополненных метаданных. В полученных таким образом результатах исключаются случайные совпадения, что особенно актуально при поиске по аббревиатурам. Набор метаданных, представленных на сайте ROR, позволяет получить ссылки на такие семантические сети, как Wikidata.

В разделе 1 представлены особенности обработки аффилиации научных документов. В разделе 2 предложен метод пополнения метаданных, представляющих аффилиацию авторов публикаций и их нормализацию в формате библиотеки Lobachevskii-DML. В разделе 3 предложен разработанный алгоритм дополнения метаданных аффилиации с помощью REST API – открытого реестра идентификаторов ROR. Алгоритм реализован на коллекции журнала «Электронные библиотеки» за 2021–2022 годы. Разработанный метод позволяет расширить набор средств для улучшения метаданных по пополнению аффилиации, приведенных в [17]. В разделе 4 предложен метод пополнения метаданных цифровой библиотеки Lobachevskii-DML.

1. ОСОБЕННОСТИ ОБРАБОТКИ АФИИЛИАЦИИ НАУЧНЫХ ДОКУМЕНТОВ

Включение метаданных аффилиаций авторов в основные метаданные научных коллекций является требованием международных научных журналов [15, 20]. Основные требования к составляющим аффилиации авторов научных публикаций, примеры влияния точности и полноты представленной в ней информации подробно приведены в [12, 13]. К основным составляющим аффилиации можно отнести такие метаданные, как полное название организации, полный адрес (индекс, номер дома, улица, город, регион, страна). Как правило, набор метаданных аффилиаций авторов в электронных коллекциях в зависимости от требования журнала и года, когда статья была выпущена, не соответствует правилам цифровой математической библиотеки. На рис. 1 приведены примеры аффилиаций в документах коллекций цифровой библиотеке Lobachevskii-DML.

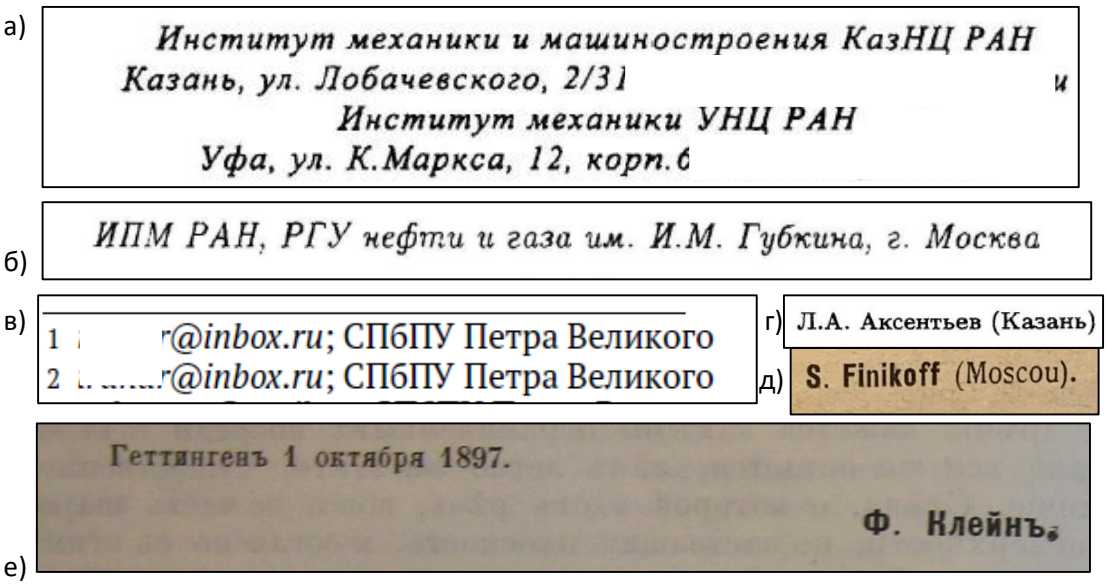

Рис. 1. Примеры аффилиаций авторов в коллекции цифровой библиотеки Lobachevskii-DML.

а) пример аффилиации, включающий название и полный адрес научной организации;

б) пример аффилиации, включающей название и город, в котором находится научная организация;

в) пример аффилиации, включающей название научной организации.

г)–е) примеры аффилиаций, включающих город, в котором работал ученый.

Приведенные выше примеры а)–г) – аффилиации статей коллекции «Труды математического центра им. Н. И. Лобачевского», д), е) – коллекции «Известия физико-математического общества при Казанском университете».

Также необходимо отметить, что набор метаданных не всегда зависит от года выпуска статьи. В архивных статьях в качестве аффилиации автора чаще всего указан только город, в котором жил автор. В данный момент аффилиация статей учитывается при подсчете показателей продуктивности научной организации, что увеличило набор метаданных и важность указания полной аффилиации в документах цифровых коллекций.

При формировании метаданных цифровых коллекций вначале необходимо определить составляющие аффилиации, при этом такие составляющие могут различаться по составу, в зависимости от требований научных изданий. Типичный состав аффилиации в различных типах коллекций предложен в таблице 1.

Таблица 1: Основные составляющие аффилиации документов коллекций цифровой библиотеки Lobachevskii-DML.

Основные составляющие аффилиации	Архивные ретро-коллекции (Коллекция «Известия физико-математического общества при Казанском университете»)	Коллекции на русском языке, выпущенные в доцифровой период (Коллекция «Труды математического центра им. Н. И. Лобачевского»)	Статьи, выпущенные в цифровую эпоху (Коллекция статей журнала «Электронные библиотеки»)
Факультет/внутренние составляющие организации	–	±	±
Организация	±	±	+
Улица, дом	–	±	±
Город, индекс	+	±	±
Страна	±	±	±

Основными проблемами при распознавании метаданных аффилиации являются: избыточные данные (например, указание кафедры университета), сокращения названий организаций, аббревиатуры, неполный набор метаданных. Также необходимо отметить, что правильный порядок указания метаданных в аффилиации является одним из основных требований в процессе автоматической обработки [13]. Также одной из задач является выделение блока аффилиации из текста статьи. Решение этой задачи сопровождается реализацией методов выделения структуры документов (см. [21–23]).

2. ОТКРЫТЫЙ РЕЕСТР ИДЕНТИФИКАТОРОВ ROR КАК ИСТОЧНИК МЕТАДАННЫХ В АФФИЛИАЦИИ АВТОРОВ

В настоящей работе использован открытый реестр идентификаторов ROR. Этот ресурс является открытым реестром идентификаторов и метаданных научных организаций по всему миру. ROR используется в системах публикации журналов, репозиториях данных, платформах управления спонсорами и грантами, рабочих процессах с открытым доступом и других компонентах исследовательской инфраструктуры для устранения неоднозначности институциональной принадлежности, улучшения обнаружения и отслеживания результатов исследований по принадлежности, а также облегчения рабочих процессов публикации открытого доступа.

Дополнение метаданных цифровых коллекций с помощью ROR может сопровождаться некоторыми особенностями.

Основные метаданные, которые мы можем извлекать из ROR:

- id – идентификатор и ссылка в системе ROR;
- name – официальное название организации;
- aliases – альтернативные названия (в случае Казанского федерального университета – «Казанский университет», «Kazan State University»);
- acronyms – сокращения (в случае Казанского федерального университета – «KFU»);
- label – название на региональном языке, а также язык, на котором приведено название;
- wikipedia_url – страница в Wikipedia;

- addresses, country, city – адрес, страна, город;
- links – ссылка на сайт организации;
- Wikidata – идентификатор в Wikidata.

Одна из главных особенностей ROR – наличие средств поиска научной организации, что позволяет достаточно точно находить по названию научной организации ее профиль. Это дает преимущество в поиске в сравнении с методами поиска по семантической сети. В частности, в алгоритмах поиска по семантической сети используется ограничение множества сущностей, что не всегда позволяет однозначно определить искомую сущность [6]. Существование в ROR Wikidata id помогает дополнить метаданные аффилиации информацией из Wikidata. В качестве ограничений использования ROR можно указать неполноту коллекции метаданных (особенно научных организаций – некоторые научные организации изменяют названия, что усложняет поиск), неполноту акронимов и альтернативных названий организаций.

Необходимость использования такого ресурса, как ROR, обусловлена тем, что в более ранних коллекциях цифровой библиотеки Lobachevskii-DML и журнала «Электронные библиотеки (rdl-journal.ru) аффилиация не является полной, что не соответствует набору основных метаданных цифровых коллекций. Наличие поискового движка, связь с семантическими сетями позволяют использовать ROR в качестве валидного источника метаданных для цифровых коллекций.

3. ДОПОЛНЕНИЕ МЕТАДАННЫХ СРЕДСТВАМИ REST API

Далее представлен алгоритм обращения к ROR. Доступ к ROR осуществляется с помощью средства доступа REST API [24]. Доступ к ресурсу ограничен –2000 запросов в 5 минут, что вполне подходит к размеру коллекции «Электронные библиотеки» за 2021–2022 годы. Также в данный момент REST API приводит только активные организации, что ограничивает набор коллекций, к которым мы можем применить алгоритм.

Доступ к REST API осуществляется средствами cURL – служебной программы командной строки (https://curl.se/).

Приведем основной алгоритм извлечения информации из ROR:

1) формирование cURL запроса к REST API;

2) получение JSON ответа на запрос для научных организаций;

3) разбор JSON файлов;

3.1) отбор результатов запроса;

3.2) перевод результатов запроса в XML.

Отметим, что при формировании запроса на шаге 1 используются сервисы метаданных, обеспечивающие извлечение блока аффилиации из документа с последующим разбором аффилиации на составляющие. Эти сервисы основаны на методах анализа структурных составляющих научных документах (см. [15]).

Для реализации данного алгоритма использованы средства языка C#, расширение Newtonsoft.Json (https://www.newtonsoft.com/json) для работы с JSON, а также System.Xml, System.Xml.Linq для работы с xml-документами.

Алгоритм протестирован на коллекции статей журнала «Электронные библиотеки» за 2021–2022 годы. В метаданных статей приведены полные названия организаций на русском и английском языках, однако не приведены адреса – обязательная часть метаданных [12]. Метаданные сформированы в формате Articulus, принятом для загрузки метаданных в научную электронную библиотеку eLIBRARY.RU [9]. Формирование метаописания сборников статей в формате Articulus описано в [14, 15].

Алгоритм 1: Получение информации об организации средствами REST API и дополнение метаданных цифровой коллекции статей журнала «Электронные библиотеки».

```
1: load XDocument EB_xml EB_Articulus.xml
//формируем список организаций
2: Set Uni;// Множество организаций
3: for each issue in EB_xml:
4: ˚˚˚for each article in issue:
5: ˚˚˚˚˚˚for each author in article:
6: ˚˚˚˚˚˚˚˚˚Uni.Add(author.orgName);// добавление организации в множество организаций
7: ˚˚˚˚˚˚end for
8: ˚˚˚end for
9: end for
```

10: if (Uni.Length >= 1)

11: {

12: ˚˚˚System.Diagnostics.Process.Start("cmd.exe", @"/C cd ""C:\Users\ ˚˚˚LJM\vcpkg\"""); //запуск командной строки

13: ˚˚˚for each U in Uni:

14: ˚˚˚˚˚˚System.Diagnostics.Process.Start("cmd.exe", @"/C curl ˚˚˚˚˚˚https://api.ror.org/organizations?query.advanced=name:" + ˚˚˚˚˚˚UNorm(U) + " > C:\\lin \\ROR\\res\\" + U + ".txt"); //запрос к ˚˚˚˚˚˚REST API и сохранение файла с ответом в папку. UNorm – ˚˚˚˚˚˚функция, которая представляет нормированное название организации

15: string[] dirs = Directory.GetFiles(path, "*.txt"); //получаем все файлы из папки с запросами

16: list Nodes;// список xml-узлов с нормализованными метаданными

17: for each name in dirs:

18: ˚˚˚jsonValue = sr.ReadLine(); //считываем файл JSON

19: ˚˚˚jsonValueN = Normal(jsonValue);//отбор организации из JSON ˚˚˚файла

20: ˚˚˚XmlDocument element = ˚˚˚JsonConvert.DeserializeXmlNode(jsonValueN);// переводим из ˚˚˚JSON в xml с помощью Newtonsoft.Json;

21: ˚˚˚XmlNode node = Normalization(element);// функция отбора ˚˚˚необходимых метаданных из xml-файла, формирование узла ˚˚˚для вставки в xml-документ.

22: ˚˚˚Nodes.Add(node);

23: end for;

24: for each issue in EB_xml:

25:˚˚˚for each article in issue:

26: ˚˚˚˚˚˚for each author in article:

27: ˚˚˚˚˚˚˚˚˚author.Add(FindOrg(Nodes));// добавление дополнительных ˚˚˚˚˚˚˚˚˚метаданных организации в xml файл

28: ˚˚˚˚˚˚end for

29: °°°end for

30: end for

31: write EB_xml in EB_Articulus_Sup.xml

32: save EB_Articulus_Sup.xml

На Рис. 2 приведен фрагмент результата запроса.


```
{ "members": [
    {
      "id":"https://ror.org/05256ym39",
    "name":"Kazan Federal University",
    "email_address":"",
    "ip_addresses":[],
    "established":1804,
    "types":["Education"],
    "relationships":[],
    "name": "Molecule Man",
      "addresses":[{
        "lat":55.78874,
        "lng":49.12214,
        "state":null,
        "state_code":null,
        "city":"Kazan'",
        "geonames_city":{
          "id":551487,
          "city":"Kazan'",
          "geonames_admin1":{
            "name":"Tatarstan Republic",
            "id":484048,
            "ascii_name":"Tatarstan Republic",
            "code":"RU.73"},
          "geonames_admin2":{
            "name":"Gorod Kazan'",
            "id":862913,
            "ascii_name":"Gorod Kazan'",
            "code":"RU.73.862913"},
          "license":{
            "attribution":"Data from geonames.org under a CC-BY 3.0 license",
            "license":"http://creativecommons.org/licenses/by/3.0/"},
          "nuts_level1":{
            "name":null,
            "code":null},
          "nuts_level2":{
            "name":null,
            "code":null},
```


Рис. 2. Фрагмент результата запроса к ROR.

Отметим, что при получении Wikidata id можно использовать алгоритм, приведенный в статье [6], что позволяет дополнять метаданные еще большим набором метаданных средствами семантической сети Wikidata. Приложение можно подключить в функции Normalization(element) (строка 22 алгоритма 1).

В ходе тестирования процент найденных в ROR аффилиаций научных организаций из данной коллекции составил 82%.

В процессе регистрации статьи в информационной системе OJS в системе Open Journal System (OJS) авторы вводят аффилиацию самостоятельно, однако она может быть неполной, например, могут отсутствовать город, страна, или аф-

филиация может быть написана в сокращенном виде. Набор метаданных OJS версии 3.0 не подразумевает ROR-идентификатора. Таким образом, мы можем хранить его во внутренних форматах (например, при формировании метаданных цифровых коллекций Lobachevskii-DML).

В дальнейшем предполагается с помощью разработанного инструмента произвести пополнение и уточнение метаданных электронных коллекций, входящих в Lobachevskii-DML. К ограничениям метода можно отнести: неполноту информации в ROR, отсутствие архивных организаций (что ограничивает использование метода в ретро-коллекциях).

4. ФОРМИРОВАНИЕ КОЛЛЕКЦИЙ ЦИФРОВОЙ МАТЕМАТИЧЕСКОЙ БИБЛИОТЕКИ LOBACHEVSKII-DML

Изучая метаданные статей электронных коллекций, можно заметить, что чаще всего для их описания используются древовидные структуры [9–11, 25]. Однако архитектура метаданных библиотеки Lobachevskii-DML имеет реляционную структуру [14]. Возникает задача преобразования коллекций в формат, пригодный для загрузки в Lobachesvkii-DML. При решении этой задачи необходимо создавать методы преобразования метаданных из xml-формата в формат MySQL.

Обозначим метаданные, приведенные на рис. 3:

- articles: включает в себя идентификатор документа, идентификатор издания, номер тома, положение статьи в номере, номера страниц, дата опубликования. Следующие таблицы включают в себя “id” как первичный ключ реляционной таблицы;
- article_files: включает ключ “id”, url ссылку на статью, а также информацию, содержится данный файл на сайте библиотеки или ссылка является внешней;
- article_ids: содержит ссылки на документ в других интернет-источниках, а также название этих источников;
- article_titles: названия документов с указанием языка, на котором представлен документ;
- article_related: 5 схожих документов внутри коллекции Lobachevskii-DML вместе со ссылкой;

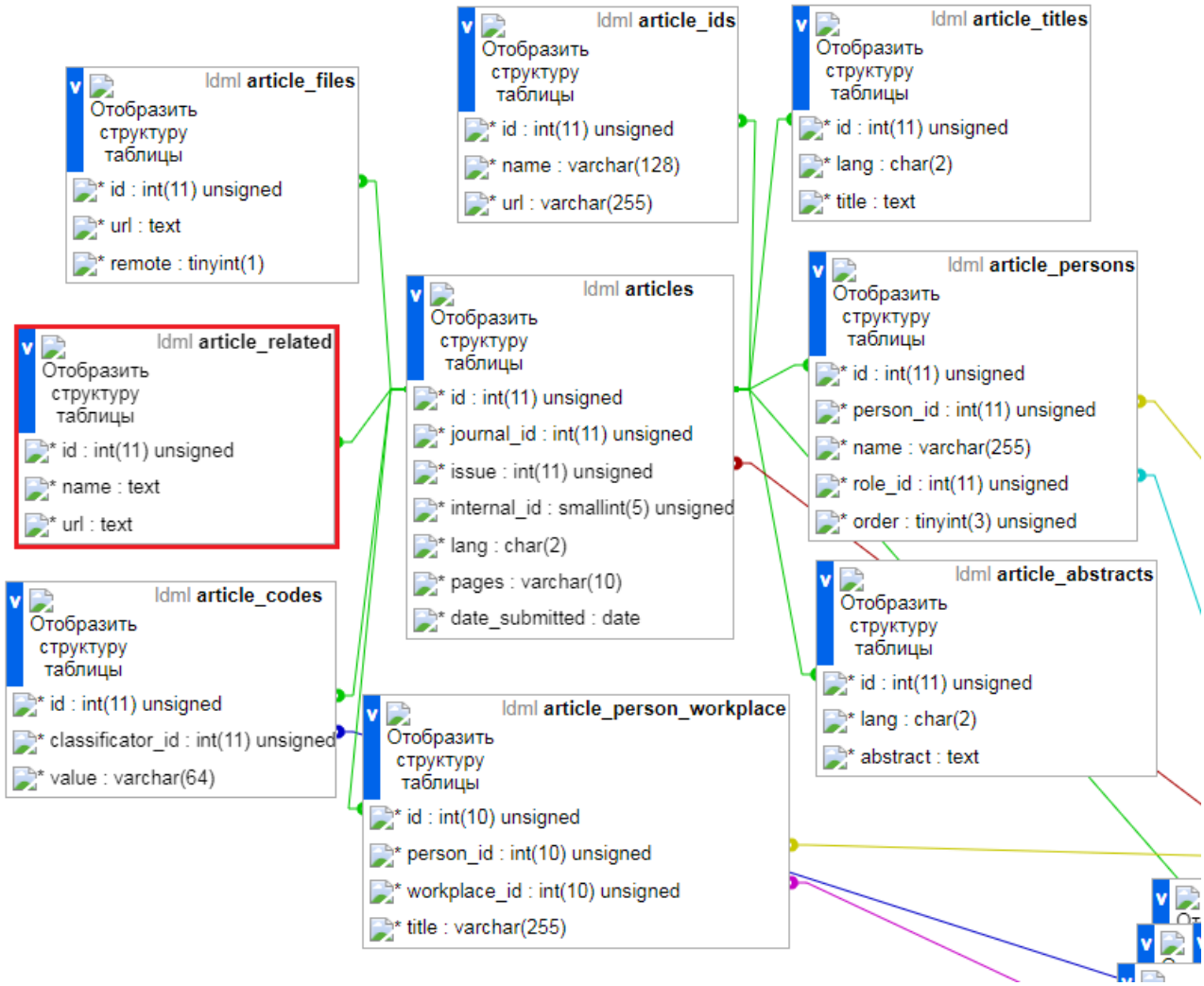

Рис. 3. Фрагмент схемы описания статьи в цифровой библиотеке Lobachevskii-DML.

- article_persons: идентификатор автора документа, имя, роль, порядок авторов в документе;
- article_codes: классификаторы и их значения;
- article_abstracts: аннотации и язык, на котором написана аннотация;
- article_person_workplace: место работы автора документа, идентификатор автора и места работы и название.

Метаданные коллекций для загрузки в цифровую библиотеку могут быть получены в различных форматах. Так как часто метаданные уже представлены в xml и форматах, основанных на xml, то необходимо создание средств преобразования коллекций в формат библиотеки Lobachevskii-DML.

Для решения этой проблемы было написано приложение, которое позволило генерировать таблицы метаданных для загрузки в цифровую библиотеку. При реализации данной задачи было необходимо описать классы, в которых содержится информация о метаданных статьи и автора, поля этих классов приведены на рис. 4.

```
class Article
{
    int id;
    int journal_id = 14;
    string volume;
    string number;
    int issue_id;
    int in_issue_id;
    string pages;
    string title_rus;
    string abstract_rus;
    string link;
    string lang = "ru";
    string date = "";
    List<string> keywords = new List<string>();
    List<Person> authors = new List<Person>();
```

```
class Person {
    int person_id;
    string surname;
    string initials;
    string fullname;
    string affiliation;
    int role_id = 1;
    int number_id=1;
    int affiliation_id;
```

Рис. 4. Поля классов, образующих метаописание статьи.

Далее приведен алгоритм перевода метаданных журнала из формата Articulus в формат, подходящий для загрузки в Lobachevskii-DML [9]. Основной спецификой алгоритма был широкий набор метаданных в метаописании статей в формате Articulus. Также проблему вызывал тот факт, что необходимо связывать уже существующие в библиотеке Lobachevskii-DML метаданные, такие, как персоналии авторов и организации, в которых они работали.

Алгоритм 2: **Алгоритм нормализации метаданных статей журнала в формат, предназначенный для загрузки в базу Lobachevskii-DML**

```
1:    load Dictionary<string, int> affiliation // создание словаря организаций, которые уже существуют в контексте Lobachvevskii-DML
2:    load Dictionary<string, int> names // создание словаря организаций, которые уже существуют в контексте Lobachvevskii-DML
3:    load files//загрузка файлов метаописания в формате xml
4:    for each file in files:
5:      volume = FindNode(“volume”)//поиск по xml дереву тома выпуска
6:      number = FindNode(“number”) // поиск по xml дереву номера выпуска
```

```
7:      id_issue = IssueIdDefiner(volume, number) // определение id номера для Lobachevskii-DML
8:      for each article in file
9:        newArticle = new Article(volume, number) // создание экземпляра класса Article
10:       newArticle.pages = FindNode("pages") // поиск номеров страниц статьи
11:       newArticle.artTitles = FindNode("artTitles", lang= "RUS") // поиск названия статьи на русском языке
12:       newArticle.abstracts = FindNode("abstracts", lang="RUS") // поиск аннотации на русском языке
13:       newArticle.keywords = FindNode("keywords", lang= "RUS") // поиск и формирование списка ключевых слов на русском языке
14:       newArticle.files = FindNode("files", lang= "RUS")
15:       newArticle.dates = FindNode("dates", lang= "RUS")
16:       for each author in paper:
17:         Person newauthor = new Person()
18:         newauthor.surname = FindNode("surname", lang="RUS") // поиск фамилии на русском языке
19:         newauthor.initials = FindNode("initials ", lang= "RUS") // поиск инициалов на русском языке
20:         newauthor.orgName = FindNode("orgName", lang= "RUS") поиск названия на русском языке
21:         newauthor.Form_fullname()//формирование полного имени
22:         affiliation = newauthor.Form_aff_id(affiliation) // добавление в словарь организации новой организации или получение id из словаря организаций
23:         names = newauthor.Form_name_id(names) // добавление в словарь имен авторов нового автора или получение id из словаря авторов
24:         newArticle.authors.Add(newauthor) // добавление нового автора в статью
25:       end for
26:       newArticle.Form() // формирование записей метаописания статьи в различные таблицы
27:     end for
28:   end for
```

При применении разработанного метода были сформированы следующие коллекции: «Метаданные статей журнала «Электронные библиотеки» за 1998–2022 гг., которая была загружена в Lobachevskii-DML, и электронная коллекция «Метаданные тезисов докладов XI Всероссийского съезда по фундаментальным проблемам теоретической и прикладной механики», также представленная в цифровой библиотеке Lobachevskii-DML [26, 27].

В рамках формирования коллекций Lobachevskii-DML возникла необходимость формирования методов загрузки дополнительных метаданных к документам цифровых коллекций [28–30]. Так, был разработан метод добавления новых метаданных к статьям, которые уже находятся в коллекции. Схема цифровой библиотеки (рис. 4) была дополнена таблицей article_related, включающей в себя наборы схожих по тематике статей к статьям, содержащимся в библиотеке. Указанный метод реализован в виде программного инструмента, размещенного в цифровой библиотеке [31].

ЗАКЛЮЧЕНИЕ

Сформирован алгоритм дополнения цифровых коллекций средствами платформы ROR, он был реализован на языке C#. Средствами этого алгоритма пополнена коллекция метаданных журнала «Электронные библиотеки» за 2021–2022 годы, сформировано приложение, которое позволяет дополнять метаданные аффилиации цифровых коллекций. В дальнейшем данный алгоритм будет применен на коллекции цифровой библиотеки Lobachevskii-DML и включен в цифровой сервис формирования метаданных «Фабрика метаданных» цифровой библиотеки Lobachevskii-DML [22].

Благодарности

Работа выполнена в КазО МСЦ РАН – филиале ФГУ ФНЦ НИИСИ РАН в рамках государственного задания FNEF-2022-0014.

СПИСОК ЛИТЕРАТУРЫ

1. *Elizarov A.M., Lipachev E.K*. Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // International Conference on Data Analytics and Management in Data Intensive Domains. 2017. P. 326–333.

2. *Елизаров А.М., Липачёв 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

3. *Elizarov A., Lipachev E.* Big math methods in Lobachevskii-DML digital library // CEUR Workshop Proceedings 2019. V. 2523. P.59–72.

4. *Carette J., Farmer W.M., Kohlhase M., Rabe F.* Big Math and the OneBrain Barrier. A Position Paper and Architecture Proposal // arXiv:1904.10405v1. 2019. https://doi.org/10.48550/arXiv.1904.10405

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

6. *Гафурова П.О., Липачёв Е.К.* Метод уточнения аффилиации авторов научных документов на основе запросов к семантической сети // Научный сервис в сети Интернет: труды XXIV Всероссийской научной конференции (19–22 сентября 2022 г., онлайн). М.: ИПМ им. М.В. Келдыша. 2022. С. 115–127. https://doi.org/10.20948/abrau-2022-31

7. *Гафурова П.О.* Гармонизация метаданных цифровых математических коллекций // Информационные технологии в образовании и науке (ИТОН-2023): материалы IX Международной научно-практической конференции в рамках IV Международного форума по математическому образованию (27 марта – 1 апреля 2023 г.) / отв. ред. А.А. Агафонов. Казань: Изд-во Академии наук РТ. 2023. С. 46–50. URL: https://kpfu.ru/portal/docs/F357733059/ITON_2023.pdf

8. *Elizarov A., Lipachev E.* Digital Libraries and the Common Digital Space of Mathematical Knowledge // CEUR Workshop Proc. 2021. V. 2990. P. 25–38.

9. Инструкция по Articulus (для периодических и непериодических изданий). URL: https://vniigis.ru/1_dlya_failov/Help/Инструкция%20по%20работе%20с%20программой%20Articulus%20eLibrary%20НЭБ%20РИНЦ.pdf

10. *Bouche T., Goutorbe C., Jorda J.-P., Jost M.* The EuDML Metadata Schema: Version 1.0. // Towards a Digital Mathematics Library, July 2011, Bertinoro, Italy. P. 45–61. URL: https://hal.univ-grenoble-alpes.fr/hal-03765892/file/D3.6.pdf

11. How can I submit metadata for a complete journal or conference? URL: https://dblp.org/faq/How+can+I+submit+meta+data+for+a+complete+journal+or+conference.html

12. *Кириллова О.В.* Аффилиация авторов научных публикаций и ее представление в статьях и в глобальных индексах цитирования. URL: https://kai.ru/documents/1489522/1535688/affiliation.pdf/a3349af1-1b8d-4f05-ba54-812f60a32e21

13. *Кириллова О.В.* Значение и основные требования к представлению аффилиации авторов в научных публикациях // Научный редактор и издатель. 2016. Т. 1 (1–4). С. 32–42.

14. *Елизаров А.М., Липачев Е.К., Хайдаров Ш.М*. Цифровая математическая библиотека Lobachevskii DML. Свидетельство о государственной регистрации базы данных № 2021620324 от 25 февраля 2021 года.

15. *Елизаров А.М., Зайцева Н.В., Зуев Д.С., Липачёв Е.К., Хайдаров Ш.М.* Сервисы формирования метаданных цифровых документов в форматах международных наукометрических баз данных // Научный сервис в сети Интернет: труды XX Всероссийской научной конференции (17–22 сентября 2018 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2018. С. 175–185. https://doi.org/10.20948/abrau-2018-53/2020610082.pdf

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

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

18. *Апанович З.В.* Информация о российских научных организациях в международных и русскоязычных источниках данных // Электронные библиотеки. 2021. Т. 24 (5). С. 756–769. URL: https://rdl-journal.ru/article/view/701

19. ROR – The Research Organization Registry (ROR). URL: https://ror.org/

20. *Елизаров А.М., Липачёв Е.К., Хайдаров Ш.М.* Программа автоматизированного формирования выпусков журнала «Электронные библиотеки» Свидетельство о государственной регистрации базы данных № 2020610082 от 9 января 2020 года.

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

22. *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

23. *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

24. ROR REST API Guide URL: https://ror.readme.io/docs/rest-api

25. Journal Archiving and Interchange Tag Library NISO JATS Version 1.3d1. URL: https://jats.nlm.nih.gov/archiving/tag-library/1.3d1/chapter/how-to-read.html

26. Электронная коллекция статей журнала «Электронные библиотеки» URL: https://lobachevskii-dml.ru/journal/elbib

27. Электронная коллекция «XI Всероссийский съезд по фундаментальным проблемам теоретической и прикладной механики».
URL: https://lobachevskii-dml.ru/conference/congress_11

28. *Гафурова П.О.* Дополнение метаданных документов цифровых коллекций из внешних источников // Материалы Всероссийской школы-конференции «Лобачевские чтения – 2023». Казань: Изд-во КФУ. 2023. Т. 67 С. 33–36.

29. *Гафурова П.О.* Автоматическое пополнение метаданных цифровых публикаций с использованием семантических сервисов сети Интернет // Научный сервис в сети Интернет. 2023. № 25. С. 84–93.
https://doi.org/10.20948/abrau-2023-27

30. *Elizarov A., Gafurova P., Lipachev E.* Algorithms for Integration of Unstructured Mathematical Documents into the Common Digital Space of Scientific Knowledge// CEUR Workshop Proc. 2021. V. 2990. P. 39–49.
URL: https://ceur-ws.org/Vol-2990/rpaper4.pdf

31. *Гафурова П.О., Кривцова В.А.* Программа формирования списка близких статей цифровой математической библиотеки на основе статистических метрик. Свидетельство о государственной регистрации базы данных № 2023684278 от 15 ноября 2023 года.

AUTOMATIC REPLENISHMENT OF METADATA OF DIGITAL PUBLICATIONS USING SEMANTIC SERVICES OF THE INTERNET

P. O. Gafurova[0000-0002-1544-155X]

National Research Centre "Kurchatov Institute"

pogafurova@gmail.com

Abstract

The article describes approaches to replenishing metadata of documents in electronic collections of a digital mathematical library. An open resource of the semantic network is used as a replenishment. For this purpose, software tools have been developed to search for the necessary data and include it in a metadata set. A separate block of metadata in a scientific article is formed from the affiliation of the authors presented in the document. Typically, the ownership that occurs in a document does not contain sufficient data to generate a set of metadata. A method has been developed for providing author affiliation metadata, providing an open register of scientific organization identifiers (ROR), as well as means for making connections between ROR and other semantic chains. This method was applied to the collections of articles of the journal "Digital Libraries" for 2021–2022.

The article describes a method for connecting the Lobachevsky digital mathematical library-DML to new electronic collections, and describes a method for transforming metadata into a digital format available for downloading.

Keywords*: ROR, Wikidata, digital libraries, affiliation metadata, Lobachevskii-DML.*

REFERENCES

1. *Elizarov A.M., Lipachev E.K*. Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // International Conference on Data Analytics and Management in Data Intensive Domains. 2017. P. 326–333.

2. *Elizarov A.M., Lipachev, E.K*. Lobachevskii digital library in the scientific space of mathematical knowledge // Nauchno-tekhnicheskaya informaciya. Seriya 1: Organizaciya i metodika informacionnoj raboty. 2023. No 1. P. 32–37. https://doi.org/10.36535/0548-0019-2023-01-3

3. *Elizarov A., Lipachev E.* Big math methods in Lobachevskii-DML digital library // CEUR Workshop Proceedings. 2019. V. 2523. P. 59–72.

4. *Carette J., Farmer W.M., Kohlhase M., Rabe F.* Big Math and the OneBrain Barrier. A Position Paper and Architecture Proposal // arXiv:1904.10405v1. 2019. https://doi.org/10.48550/arXiv.1904.10405

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

6. *Gafurova P.O., Lipachev E.K.* Method for Clarifying the Affiliation of Authors of Scientific Documents Based on Requests to the Semantic Web// Scientific Services and Internet, SSI-2022. M.: IPM im. M.V. Keldysha. 2022. P. 115–127.

7. *Gafurova P.O.* Harmonization of metadata in digital mathematical libraries // Informacionnye tekhnologii v obrazovanii i nauke (ITON-2023): materialy IX Mezhdunarodnoj nauchno-prakticheskoj konferencii v ramkah IV Mezhdunarodnogo foruma po matematicheskomu obrazovaniyu (27 marta – 1 aprelya 2023 g.) / otv. red. A.A. Agafonov. Kazan': Izd-vo Akademii nauk RT. 2023. P. 46–50. URL: https://kpfu.ru/portal/docs/F357733059/ITON_2023.pdf

8. *Elizarov A., Lipachev E.* Digital Libraries and the Common Digital Space of Mathematical Knowledge // CEUR Workshop Proc. 2021. V. 2990. P. 25–38.

9. Articulus user manual (for periodicals and nonperiodicals) URL: https: //vniigis.ru/1_dlya_failov/Help/Инструкция%20по%20работе%20с%20программой%20Articulus%20eLibrary%20НЭБ%20РИНЦ.pdf

10. *Bouche T., Goutorbe C., Jorda J.-P., Jost M.* The EuDML Metadata Schema: Version 1.0. Towards a Digital Mathematics Library, July 2011, Bertinoro, Italy. P.45–61. https://hal.univ-grenoble-alpes.fr/hal-03765892/file/D3.6.pdf

11. How can I submit metadata for a complete journal or conference? https://dblp.org/faq/How+can+I+submit+meta+data+for+a+complete+journal+or+conference.html

12. *Kirillova O.V.* Affiliaciya avtorov nauchnyh publikacij i ee predstavlenie v stat'yah i v global'nyh indeksah citirovaniya.
URL: https://kai.ru/documents/1489522/1535688/affiliation.pdf/a3349af1-1b8d-4f05-ba54-812f60a32e21

13. *Kirillova O.V.* Znachenie i osnovnye trebovaniya k predstavleniyu affiliacii avtorov v nauchnyh publikaciyah // Nauchnyj redaktor i izdatel'. 2016. V. 1 (1–4). P. 32–42.

14. *Elizarov A.M., Lipachev E.K., Khajdarov Sh.M.* Cifrovaya matematicheskaya biblioteka Lobachevskii DML. Svidetel'stvo o gosudarstvennoj registracii bazy dannyh № 2021620324 ot 25 fevralya 2021 goda.

15. *Elizarov A.M., Zaitseva N., Zuev D.S., Lipachev E.K., Khajdarov Sh.M.* Services for Formation of Digital Documents Metadata in the Formats of International Science-Based Databases // CEUR Workshop Proc. 2018. V. 2260. P. 175–185.
URL: https://ceur-ws.org/Vol-2260/53_175-185.pdf

16. *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// Scientific Services and Internet, SSI-2021. M.: IPM im. M.V. Keldysha. 2021. P. 22–33. https://doi.org/10.20948/abrau-2021-22
URL: https://keldysh.ru/abrau/2021/theses/22.pdf

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

18. *Apanovich Z.V.* Information about russian research organizations in multilingual data sources // Russian Digital Libraries Journal. 2021. V. 24 (5). P. 756–769. URL: https://rdl-journal.ru/article/view/701

19. ROR – The Research Organization Registry (ROR) // https://ror.org/

20. *Elizarov A.M., Lipachyov E.K., Hajdarov SH.M.* Programma avtomatizirovannogo formirovaniya vypuskov zhurnala «Elektronnye biblioteki» Svidetel'stvo o gosudarstvennoj registracii bazy dannyh № 2020610082 ot 9 yanvarya 2020 goda.

21. *Gafurova P., Elizarov A., Lipachev E.* Basic services of factory metadata digital mathematical library Lobachevskii-DML // Russian Digital Libraries Journal. 2020. V. 23 (3). P. 336–381. https://doi.org/10.26907/1562-5419-2020-23-3-336-381

22. *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

23. *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

24. ROR REST API Guide URL: https://ror.readme.io/docs/rest-api

25. Journal Archiving and Interchange Tag Library NISO JATS Version 1.3d1 URL: https://jats.nlm.nih.gov/archiving/tag-library/1.3d1/chapter/how-to-read.html

26. Electronic collection of scientific journal articles «Russian Digital Libraries Journal» URL: https://lobachevskii-dml.ru/journal/elbib

27. Electronic collection of scientific journal articles «XI Vserossijskij s"ezd po fundamental'nym problemam teoreticheskoj i prikladnoj mekhaniki».
URL: https://lobachevskii-dml.ru/conference/congress_11

28. *Gafurova P.O.* Enhancing digital collections document metadata from external sources // Materialy Vserossijskoj shkoly-konferencii ”Lobachevskie chteniya – 2023” Kazan': Izd-vo KFU. 2023. V. 67 P. 33–36.

29. *Gafurova P.O.* Automatic replenishment of metadata of digital publications using semantic services of the Internet // Scientific Services and Internet, SSI-2023. M.: IPM im. M.V. Keldysha. 2023. V. 25. P. 84–93.
https://doi.org/10.20948/abrau-2023-27

30. *Elizarov A., Gafurova P., Lipachev E.* Algorithms for Integration of Unstructured Mathematical Documents into the Common Digital Space of Scientific Knowledge// CEUR Workshop Proc. 2021. V. 2990. P.39–49.
https://ceur-ws.org/Vol-2990/rpaper4.pdf

31. *Gafurova P.O., Krivcova V.A.* Programma formirovaniya spiska blizkih statej cifrovoj matematicheskoj biblioteki na osnove statisticheskih metrik. Svidetel'stvo o gosudarstvennoj registracii bazy dannyh № 2023684278 ot 15 noyabrya 2023 goda.

СВЕДЕНИЯ ОБ АВТОРЕ

ГАФУРОВА Полина Олеговна – младший научный сотрудник Казанского филиала Межведомственного суперкомпьютерного центра Российской академии наук, младший научный сотрудник Национального исследовательского центра «Курчатовский институт», Лаборатории суперкомпьютерного моделирования.

Polina Olegovna GAFUROVA – Junior Researcher, Joint Supercomputer Center of the Russian Academy of Sciences, Junior Researcher, National Research Centre “Kurchatov Institute”.

email: pogafurova@gmail.com;

ORCID: 0000-0002-1544-155X

Материал поступил в редакцию 15 января 2024 года