

# Автоматическое пополнение метаданных цифровых публикаций с использованием семантических сервисов сети Интернет

П.О. Гафурова<sup>1</sup> [0000-0002-1544-155X]

<sup>1</sup>*Казанский филиал Межведомственного суперкомпьютерного центра Российской академии наук*

**Аннотация.** В статье описаны подходы к дополнению метаданных средствами семантических сервисов сети Интернет. Разработан метод дополнения метаданных аффилиации средствами реестра ROR – открытого реестра идентификаторов научных организаций и средства осуществления связей между ROR и другими семантическими сетями. Этот метод реализован на коллекции статей журнала «Электронные библиотеки» за 2021-2022 годы.

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

## Automatic replenishment of metadata of digital publications using semantic services of the Internet

P.O. Gafurova<sup>1</sup> [0000-0002-1544-155X]

<sup>1</sup>*Joint Supercomputer Center of the Russian, Academy of Sciences, Kazan, Russian Federation*

**Abstract.** The article describes approaches to supplementing metadata by means of semantic networks. A method has been developed for supplementing affiliation metadata with the help of the ROR registry - an open registry of identifiers of scientific organizations and a means of making links between ROR and other semantic networks.

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

## 1. Введение

Одной из задач, решаемых в рамках проекта создания цифровой математической библиотеки Lobachevskii-DML (<https://lobachevskii-dml.ru/>), является разработка методов формирования метаданных документов электронных коллекций. В частности, ставятся задачи пополнения метаданных электронных коллекций и задачи гармонизации коллекций метаданных [1–3]. Под гармонизацией метаданных мы понимаем дополнение метаданными архивных коллекций Lobachevskii-DML.

Существенными, но отсутствующими метаданными являются метаданные аффилиации. Особенно эта проблема заметна при формировании ретро коллекции: в статье указывается город в котором написана статья, без указания организации.

Основные требования к составляющим аффилиации авторов научных публикаций, примеры влияния точности и полноты представленной в ней информации, приведены в [4, 5].

При формировании метаданных электронных коллекций цифровой библиотеки Lobachevskii-DML задача автоматического извлечения и пополнения составляющих аффилиации авторов публикаций является одной из наиболее сложных. В существенной части документов электронных коллекций присутствуют минимальные сведения об организации, что не позволяет без дополнительных сервисов и ручной обработки составить полную аффилиацию авторов документов. Например, в ретро-коллекциях в статьях указан только город, в котором работал автор, причем у самих авторов указана только фамилия с одним инициалом, что затрудняет подготовку метаданных аффилиации [6, 7].

Пополнение метаданных цифровых коллекций можно осуществлять различными способами. Одним из таких способов является использование семантических сетей в сети Интернет (примеры использования таких сетей приведены в [1, 6, 7]). Особенностью использования семантических сетей, таких как Wikidata (<https://www.wikidata.org/>) или DBpedia (<https://www.dbpedia.org/>) является формирование запросов поиска по семантической сети [1, 7]. Другим подходом к дополнению метаданных является поиск в специализируемых семантических ресурсах, в данной статье в качестве этого ресурса мы рассматриваем открытый реестр идентификаторов ROR (The Research Organization Registry) (<https://ror.org/>) [8]. Использование специализированных коллекций улучшает качество дополненных метаданных, так как в результатах не попадают случайные совпадения, что особенно актуально при поиске по аббревиатурам, а также специфика метаданных, которые представлены на сайте ROR, позволяет получить ссылки на такие семантические сети как Wikidata.

В работе представлен метод дополнения метаданных, представляющих аффилиацию авторов публикаций и их нормализацию в формате Lobachevskii-DML. Разработан 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, приведенные в [7].

## **2. Открытый реестр идентификаторов ROR как источник метаданных в аффилиации авторов**

В настоящей работе используется открытый реестр идентификаторов ROR. Он является открытым реестром идентификаторов и метаданных университетов 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. Так основные метаданные, которые мы можем использовать 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 университета;
- `Wikidata` – идентификатор в Wikidata.

Одна из главных особенностей ROR – это поисковой движок, который позволяет достаточно точно находить по названию научной организации ее профиль. Это дает преимущество в поиске в сравнении с поиском по семантической сети. В работе [1] отмечено, что поиск страницы сущности по Wikidata – это ограничение множества всех сущностей Wikidata с помощью, например, города, что не всегда возможно. 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 меняют названия, что усложняет поиск), неполноту акронимов и альтернативных названий организаций.

Необходимость использования такого ресурса как ROR обусловлена тем, что в более ранних коллекциях Lobachevskii-DML и журнала Электронные библиотеки аффилиация была не полной, что не соответствует набору основных метаданных цифровых коллекций (набор основных метаданных приведен в [4, 5]). Наличие поискового движка, связь с семантическими сетями позволяет использовать ROR в качестве валидного источником метаданных для цифровых коллекций.

### **3. Дополнение метаданных средствами REST API**

Далее представлен алгоритм обращения к ROR. Доступ к ROR осуществляется посредством REST API (<https://ror.readme.io/docs/rest-api>). Ограничения – 2000 запросов в 5 минут, что вполне подходит к размеру нашей коллекции. Также, в данный момент REST API приводит только активные организации, что ограничивает набор коллекций, к которым мы можем применять алгоритм.

Мы получаем доступ к REST API с помощью средства cURL (<https://curl.se/>). cURL используется при работе с командными строками.

Приведем основной алгоритм доступа к ROR:

- 1) формирование cURL запроса к REST API;
- 2) получение JSON ответа на запрос для университета;
- 3) разбор JSON файлов;
  - 3.1) отбор результатов запроса;
  - 3.2) перевод результатов запроса в XML.

Для реализации данного алгоритма используются 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 был протестирован на коллекции статей «Электронные библиотеки» за 2021–2022 годы. В метаданных статей приведены полные названия организаций на русском и английском языке, однако не приведены адреса – обязательная часть метаданных [5]. Метаданные сформированы в формате Articulus – формате, принятом для загрузки метаданных в eLIBRARY.RU – научную электронную библиотеку [9]. Формирование метаописания сборников статей в формате Articulus описано в статьях [10, 11].

---

**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
   ""университета
15: }
16: string[] dirs = Directory.GetFiles(path, "*.txt"); //получаем все файлы
   из папки с запросами
17: list Nodes;// список xml узлов с нормализованными метаданными
18: for each name in dirs:
19:   ""jsonValue = sr.ReadLine(); //считываем файл JSON
20:   ""jsonValueN = Normal(jsonValue);//отбор организации из JSON
   ""файла
21:   ""XmlDocument   element   =
   ""JsonConvert.DeserializeXmlNode(jsonValueN);// переводим из
   ""JSON в xml с помощью Newtonsoft.Json;
22:   ""XmlNode node = Normalization(element);// функция отбора
   ""необходимых метаданных из xml файла, формирование узла
   ""для вставки в xml документ.
23:   ""Nodes.Add(node);
24: end for;
25: for each issue in EB_xml:
26:   ""for each article in issue:
27:     ""for each author in article:
28:       ""author.Add(FindOrg(Nodes));// добавление дополнительных
   ""метаданных организации в xml файл

```

---

---

```
29: °°°°° end for
30: °° end for
31: end for
32: write EB_xml in EB_Articulus_Sup.xml
33: save EB_Articulus_Sup.xml
```

---

На рис. 1 приведен фрагмент результата запроса.

```
1 { .. "members": [ CRIF
2 .. { CRIF
3 .. { "id": "https://ror.org/05256ym39", CRIF
4 .. "name": "Kazan Federal University", CRIF
5 .. "email_address": "", CRIF
6 .. "ip_addresses": [], CRIF
7 .. "established": 1804, CRIF
8 .. "types": ["Education"], CRIF
9 .. "relationships": [], CRIF
10 .. "name": "Molecule Man", CRIF
11 .. "addresses": [ { CRIF
12 .. "lat": 55.78874, CRIF
13 .. "lng": 49.12214, CRIF
14 .. "state": null, CRIF
15 .. "state_code": null, CRIF
16 .. "city": "Kazan'", CRIF
17 .. "geonames_city": { CRIF
18 .. "id": 551487, CRIF
19 .. "city": "Kazan'", CRIF
20 .. "geonames_admin1": { CRIF
21 .. "name": "Tatarstan Republic", CRIF
22 .. "id": 484048, CRIF
23 .. "ascii_name": "Tatarstan Republic", CRIF
24 .. "code": "RU.73", CRIF
25 .. "geonames_admin2": { CRIF
26 .. "name": "Gorod Kazan'", CRIF
27 .. "id": 862913, CRIF
28 .. "ascii_name": "Gorod Kazan'", CRIF
29 .. "code": "RU.73.862913", CRIF
30 .. "license": { CRIF
31 .. "attribution": "Data from geonames.org under a CC-BY 3.0 license", CRIF
32 .. "license": "http://creativecommons.org/licenses/by/3.0/", CRIF
33 .. "nuts_level1": { CRIF
34 .. "name": null, CRIF
35 .. "code": null, CRIF
36 .. "nuts_level2": { CRIF
37 .. "name": null, CRIF
38 .. "code": null, CRIF
```

Рис. 1. Фрагмент результата запроса к ROR

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

Приведённый выше алгоритм был протестирован на коллекции журнала «Электронные библиотеки» за 2021–2022 годы. В ходе тестирования процент найденных в ROR аффилиаций научных организаций из данной коллекции составил 82%.

В процессе подготовки журнала в системе OJS авторы вводят свою аффилиацию, однако она может быть неполной: может отсутствовать город, страна, или аффилиация может быть написана в сокращенном виде. Набор метаданных OJS версии 3.0 не подразумевает ROR-идентификатора. Таким

образом, мы можем хранить его в внутренних форматах (например при формировании метаописания Lobachevskii-DML).

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

#### **4. Заключение**

Был сформирован алгоритм дополнения цифровых коллекций средствами платформы ROR, он был реализован на языке C#.

Средствами этого алгоритма была пополнена коллекция метаданных журнала «Электронные библиотеки» за 2021–2022 годы, было 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 [12].

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

#### **Литература**

1. 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.
2. 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, стр. 46–50. URL: [https://kpfu.ru/portal/docs/F357733059/ITON\\_2023.pdf](https://kpfu.ru/portal/docs/F357733059/ITON_2023.pdf)
3. Elizarov A., Lipachev E. Digital Libraries and the Common Digital Space of Mathematical Knowledge // CEUR Workshop Proceedings. — 2021. — V. 2990. — P. 25–38.
4. Кириллова О.В. Аффилиация авторов научных публикаций и ее представление в статьях и в глобальных индексах цитирования —

- <https://kai.ru/documents/1489522/1535688/affiliation.pdf/a3349af1-1b8d-4f05-ba54-812f60a32e21>
5. 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.
  6. Андреичев М.Д., Гафурова П.О., Елизаров А.М., Липачёв Е.К. Пополнение метаданных документов математических цифровых ретроколлекций 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>  
<https://keldysh.ru/abrau/2021/theses/22.pdf>
  7. 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.
  8. ROR – The Research Organization Registry (ROR) // <https://ror.org/> (статья апанович в ЭБ <https://rdl-journal.ru/article/view/701>)
  9. Свидетельство о государственной регистрации программы для ЭВМ № 2023611260 Российская Федерация. Articulus : № 2023610217 : заявл. 11.01.2023 : опубл. 18.01.2023 / Д. И. Свечников ; заявитель Общество с ограниченной ответственностью Научная электронная библиотека. — EDN FEVREK.
  10. 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>
  11. Свидетельство 2020610082 Российская Федерация. Программа автоматизированного формирования выпусков журнала «Электронные библиотеки» : свидетельство об офиц. регистрации программы для ЭВМ / А.М. Елизаров, Е.К. Липачёв, Ш.М. Хайдаров; заявитель и правообладатель ФГАОУ ВО КФУ (RU). - №2020610082; заявл. 20.12.2019; опубл. 09.01.2020, Реестр программ для ЭВМ. - [1] с.
  12. 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>



## References

1. Gafurova P.O., Lipachyov E.K. Metod utochneniya affiliacii avtorov nauchnyh dokumentov na osnove zaprosov k semanticheskoj seti // Nauchnyj servis v seti Internet: trudy XXIV Vserossijskoj nauchnoj konferencii (19–22 sentyabrya 2022 g., onlajn). — M.: IPM im. M.V.Keldysha, 2022. — P. 115–127.
2. Gafurova P.O. Garmonizaciya metadannyh cifrovyyh matematicheskikh kollekcij // Informacionnye tekhnologii v obrazovanii i nauke (ITON-2023): materialy IX Mezhdunarodnoj nauchno-prakticheskoj konferencii v ramkah IV Mezhdunarodnogo foruma po matematicheskomu obrazovaniiyu (27 marta – 1 aprelya 2023 g.) / otv. red. A.A. Agafonov. – Kazan': Izd-vo Akademii nauk RT, 2023, str. 46–50. URL: [https://kpfu.ru/portal/docs/F357733059/ITON\\_2023.pdf](https://kpfu.ru/portal/docs/F357733059/ITON_2023.pdf).
3. Elizarov A., Lipachev E. Digital Libraries and the Common Digital Space of Mathematical Knowledge // CEUR Workshop Proceedings. — 2021. — V. 2990. — P. 25–38.
4. Kirillova O.V. Affiliaciya avtorov nauchnyh publikacij i ee predstavlenie v stat'yah i v global'nyh indeksah citirovaniya — <https://kai.ru/documents/1489522/1535688/affiliation.pdf/a3349af1-1b8d-4f05-ba54-812f60a32e21>
5. Kirillova O.V. Znachenie i osnovnye trebovaniya k predstavleniyu affiliacii avtorov v nauchnyh publikacijah // Nauchnyj redaktor i izdatel'. 2016. — T. 1 (1–4). — S. 32–42.
6. Andreichev M.D., Gafurova P.O., Elizarov A.M., Lipachyov E.K. Popolnenie metadannyh dokumentov matematicheskikh cifrovyyh retro-kollekcij metodom semanticheskikh setej // Nauchnyj servis v seti Internet: trudy XXIII Vserossijskoj nauchnoj konferencii (20–23 sentyabrya 2021 g., onlajn). — M.: IPM im. M.V.Keldysha, 2021. — S. 22–33. — <https://doi.org/10.20948/abrau-2021-22>  
<https://keldysh.ru/abrau/2021/theses/22.pdf>, last accessed 2022/07/07
7. 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.
8. ROR – The Research Organization Registry (ROR) // <https://ror.org/>
9. Svidetel'stvo o gosudarstvennoj registracii programmy dlja JeVM № 2023611260 Rossijskaja Federacija. Articulus : № 2023610217 : zavavl. 11.01.2023 : opubl. 18.01.2023 / D. I. Svechnikov ; zavavitel' Obshhestvo s ogranichennoj otvetstvennost'ju Nauchnaja jelektronnaja biblioteka. – EDN FEVREK.
10. Elizarov A.M., Zajceva N.V., Zuev D.S., Lipachyov E.K., Hajdarov SH.M. Servisy formirovaniya metadannyh cifrovyyh dokumentov v formatah mezhdunarodnyh naukometriceskikh baz dannyh // Nauchnyj servis v seti Internet: trudy XX Vserossijskoj nauchnoj konferencii (17-22 sentyabrya

- 2018 g., g. Novorossiysk). — M.: IPM im. M.V. Keldysha, 2018. — S. 175-185. <https://doi.org/10.20948/abrau-2018-53/2020610082.pdf>
11. Svidetel'stvo 2020610082 Rossijskaya Federaciya. Programma avtomatizirovannogo formirovaniya vypuskov zhurnala «Elektronnye biblioteki» : svidetel'stvo ob ofic. registracii programmy dlya EVM / A.M. Elizarov, E.K. Lipachyov, SH.M. Hajdarov; zayavitel' i pravoobladatel' FGAOU VO KFU (RU). - №2020610082; zayavl. 20.12.2019; opubl. 09.01.2020, Reestr programm dlya EVM. - [1] s.
  12. Gafurova P.O., Elizarov A.M., Lipachyov E.K. Bazovye servisy fabriki metadannyh cifrovoj matematicheskoj biblioteki Lobachevskii-DML // Elektronnye biblioteki. 2020. — T. 23 (3). — S. 336–381. — <https://doi.org/10.26907/1562-5419-2020-23-3-336-381>