ИПМ им.М.В.Келдыша РАН

Абрау-2023 • Труды конференции

П.О. Гафурова

Автоматическое пополнение метаданных цифровых публикаций с использованием семантических сервисов сети Интернет

Рекомендуемая форма библиографической ссылки

Гафурова П.О. Автоматическое пополнение метаданных цифровых публикаций с использованием семантических сервисов сети Интернет // Научный сервис в сети Интернет: труды XXV Всероссийской научной конференции (18-21 сентября 2023 г., онлайн). — М.: ИПМ им. М.В.Келдыша, 2023. — С. 84-93.
https://doi.org/10.20948/abrau-2023-27
https://keldysh.ru/abrau/2023/theses/27.pdf

Видеозапись выступления

Презентация к докладу

Автоматическое пополнение метаданных цифровых публикаций с использованием семантических сервисов сети Интернет

П.О. Гафурова[1] [0000-0002-1544-155X]

[1]*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[1] [0000-0002-1544-155X]

[1]*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 [1]. В частности, ставятся задачи пополнения метаданных электронных коллекций и задачи гармонизации коллекций метаданных [2–4]. Под гармонизацией метаданных мы понимаем дополнение наборов метаданных 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 [5, 6].

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 [7, 8].

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 [2, 7, 8]). Особенностью использования семантических сетей, таких как Wikidata (https://www.wikidata.org/) или DBpedia (https://www.dbpedia.org/) является формирование запросов поиска по семантической сети [2, 8]. 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/) [9, 10]. 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, позволяют 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 [8].

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 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 – 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 [2] отмечено, что поиск страницы сущности по Wikidata – это процесс ограничения множества всех сущностей Wikidata по признаку, указанному в запросе, что не всегда позволяет однозначно определить искомую сущность. 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 цифровой библиотеки Lobachevskii-DML и журнала Электронные библиотеки аффилиация была не полной, что не соответствует набору основных метаданных цифровых коллекций (набор основных метаданных приведен в [5, 6]). 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 – 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 организации;
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 [6]. Метаданные сформированы в формате Articulus, принятом для загрузки метаданных в научную электронную библиотеку eLIBRARY.RU [11]. Формирование метаописания сборников статей в формате Articulus описано в статьях [12, 13].

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

```
load XDocument EB_xml EB_Articulus.xml
//формируем список организаций
Set Uni;// Множество организаций
for each issue in EB_xml:
°°°for each article in issue:
°°°°°°for each author in article:
°°°°°°°°°Uni.Add(author.orgName);// добавление организации в множество организаций
°°°°°°end for
°°°end for
end for
if (Uni.Length >= 1)
{
°°°System.Diagnostics.Process.Start("cmd.exe", @"/C cd ""C:\Users\
°°°LJM\vcpkg\""""); //запуск командной строки
°°°for each U in Uni:
°°°°°°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
}
string[] dirs = Directory.GetFiles(path, "*.txt"); //получаем все файлы из папки с запросами
list Nodes;// список xml узлов с нормализованными метаданными
for each name in dirs:
°°°jsonValue = sr.ReadLine(); //считываем файл JSON
°°°jsonValueN = Normal(jsonValue);//отбор организации из JSON
°°°файла
°°°XmlDocument element =
°°°JsonConvert.DeserializeXmlNode(jsonValueN);// переводим из
°°°JSON в xml с помощью Newtonsoft.Json;
°°°XmlNode node = Normalization(element);// функция отбора
°°°необходимых метаданных из xml файла, формирование узла
°°°для вставки в xml документ.
°°°Nodes.Add(node);
end for;
for each issue in EB_xml:
°°°for each article in issue:
```

```
25:  ˚˚˚˚˚˚for each author in article:
26:  ˚˚˚˚˚˚˚˚˚author.Add(FindOrg(Nodes));// добавление дополнительных
27:  ˚˚˚˚˚˚˚˚˚метаданных организации в xml файл
28:  ˚˚˚˚˚˚end for
     ˚˚˚end for
29:  end for
30:  write EB_xml in EB_Articulus_Sup.xml
31:  save EB_Articulus_Sup.xml
32:
33:
```

На Рис. 1 приведен фрагмент результата запроса.

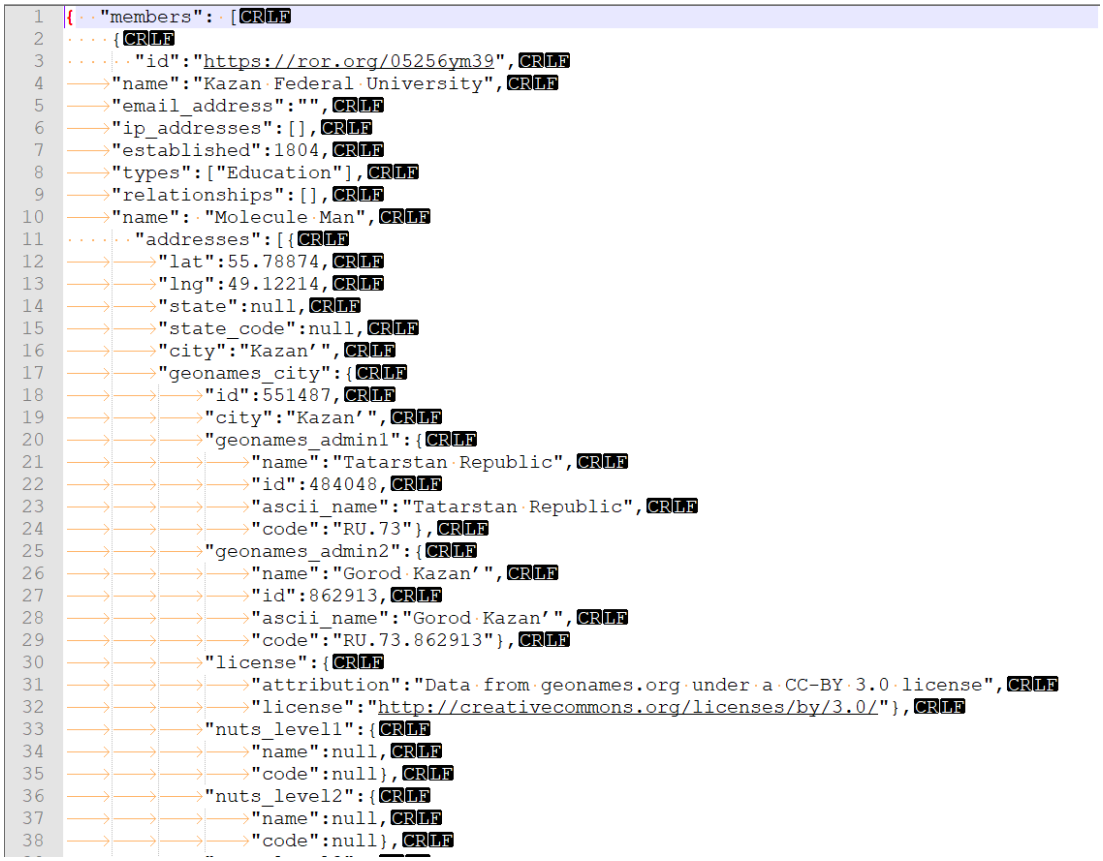

Рис. 1. Фрагмент результата запроса к ROR

Отметим, что при получении Wikidata id, мы можем использовать алгоритм, приведенный в статье [2], 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 аффилиацию самостоятельно, 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 [14].

Работа выполнена при финансовой поддержке Российского научного фонда FNEF-2022-0014.

Литература

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. 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.

3. 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
4. Elizarov A., Lipachev E. Digital Libraries and the Common Digital Space of Mathematical Knowledge // CEUR Workshop Proceedings. — 2021. — V. 2990. — P. 25–38.
5. 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
6. 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.
7. 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 https://keldysh.ru/abrau/2021/theses/22.pdf
8. 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.
9. ROR – The Research Organization Registry (ROR) // https://ror.org/
10. Апанович З. В. Информация о российских научных организациях в международных и русскоязычных источниках данных // Электронные библиотеки Т. 24 (5), 2021. С. 756–769. URL: https://rdl-journal.ru/article/view/701
11. 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.
12. 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
13. 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.
14. 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. 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. Gafurova P.O., Lipachev 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.
3. Gafurova P.O Garmonizaciya metadannyh cifrovyh matematicheskih kollekcij // 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, str. 46–50. URL: https://kpfu.ru/portal/docs/F357733059/ITON_2023.pdf.
4. Elizarov A., Lipachev E. Digital Libraries and the Common Digital Space of Mathematical Knowledge // CEUR Workshop Proceedings. — 2021. — V. 2990. — P. 25–38.
5. 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
6. Kirillova O.V. Znachenie i osnovnye trebovaniya k predstavleniyu affiliacii avtorov v nauchnyh publikaciyah // Nauchnyj redaktor i izdatel'. 2016. — T. 1 (1–4). — S. 32–42.

7. Andreichev M.D., Gafurova P.O., Elizarov A.M., Lipachev E.K. Popolnenie metadannyh dokumentov matematicheskih cifrovyh retro-kollekcij metodom semanticheskih 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
8. 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.
9. ROR – The Research Organization Registry (ROR) // https://ror.org/
10. Apanovich Z. V. Information about Russian Research Organizations in Multilingual Data Sources // RDLJ. V. 24 (5), 2021. P. 756–769. URL: https://rdl-journal.ru/article/view/701
11. Svidetel'stvo o gosudarstvennoj registracii programmy dlja JeVM № 2023611260 Rossijskaja Federacija. Articulus : № 2023610217 : zajavl. 11.01.2023 : opubl. 18.01.2023 / D. I. Svechnikov ; zajavitel' Obshhestvo s ogranichennoj otvetstvennost'ju Nauchnaja jelektronnaja biblioteka. – EDN FEVREK.
12. Elizarov A.M., Zajceva N.V., Zuev D.S., Lipachev E.K., Khajdarov SH.M. Servisy formirovaniya metadannyh cifrovyh dokumentov v formatah mezhdunarodnyh naukometricheskih baz dannyh // Nauchnyj servis v seti Internet: trudy XX Vserossijskoj nauchnoj konferencii (17-22 sentyabrya 2018 g., g. Novorossijsk). — M.: IPM im. M.V. Keldysha, 2018. — S. 175-185. https://doi.org/10.20948/abrau-2018-53/2020610082.pdf
13. 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. Lipachev, SH.M. Khajdarov; zayavitel' i pravoobladatel' FGAOU VO KFU (RU). - №2020610082; zayavl. 20.12.2019; opubl. 09.01.2020, Reestr programm dlya EVM. - [1] s.
14. Gafurova P.O., Elizarov A.M., Lipachev 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