Метод уточнения аффилиации авторов научных документов на основе запросов к семантической сети

П.О. Гафурова[1][0000-0002-1544-155X], Е.К. Липачёв[2][0000-0001-7789-2332]

[1]Казанский филиал Межведомственного суперкомпьютерного центра Российской академии наук
[2]Институт информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета

Аннотация. В соответствии с xml-схемами основных агрегирующих математических библиотек, наборы метаданных должны включать информацию о научной организации, в которой проведено исследование, представленное в статье. Эта информация в настоящее время оформляется в виде аффилиации авторов, являющейся обязательным атрибутом современной научной публикации. Состав аффилиации может быть неполным и содержать устаревшую информацию, что не позволяет сформировать требуемый набор метаданных. В настоящей работе предложены методы уточнения и пополнения метаданных аффилиации авторов документов электронных коллекций, входящих в состав цифровой математической библиотеки. В качестве источника уточнения и пополнения составляющих аффилиации авторов использованы открытые ресурсы Семантической сети. Извлечение блока аффилиации, разбор её составляющих и их анализ, выполняются с помощью программных инструментов фабрики метаданных цифровой математической библиотеки Lobachevskii DML. С помощью разработанной системы запросов к открытым источникам данных создаются семантические связи составляющих аффилиации с сетевыми информационными объектами. Это позволило на основе последующего анализа составляющих аффилиации выполнить их уточнение, а в случае отсутствия необходимых данных произвести пополнение составляющих аффилиации.

Ключевые слова: аффилиация, метаданные, семантические связи, цифровая математическая библиотека Lobachevskii DML, фабрика метаданных, электронная коллекция.

Method for Clarifying the Affiliation of Authors of Scientific Documents Based on Requests to the Semantic Web

P.O. Gafurova[1][0000-0002-1544-155X], E.K. Lipachev[2][0000-0001-7789-2332]

[1] Joint Supercomputer Center of the Russian, Academy of Sciences, Kazan, Russian Federation
[2] Institute of Information Technologies and Intelligent Systems, Kazan (Volga Region) Federal University

Abstract. In accordance with the xml-schemes of aggregating mathematical libraries, metadata sets must contain information about the scientific organization in which the study was conducted. Currently, this information is formalized as an affiliation of the authors. Affiliation is a mandatory attribute of a modern scientific publication. The affiliation may be incomplete and may contain outdated information. This does not allow to form the required set of metadata. In this paper, we propose methods for refining and replenishing the metadata of documents of electronic mathematical collections. We use the open resources of the Semantic Network as a source for refining and replenishing the authors' affiliation components. Using the software tools of the metadata factory of the Lobachevskii DML digital mathematical library, the affiliation block is extracted, its components are parsed and analyzed.Semantic links of affiliation components with network information objects are created using the system of requests to open data sources developed by us. This made it possible, on the basis of the subsequent analysis of the components of the affiliation, to clarify them. In the absence of the necessary data, the components of the affiliation are replenished.

Keywords: affiliation, metadata, semantic links, Lobachevskii Digital Mathematical Library, metadata factory, electronic collection.

1. Введение

Одной из задач, решаемых в рамках проекта создания цифровой математической библиотеки Lobachevskii DML, является разработка методов формирования метаданных документов электронных коллекций (см., например, [1]). При этом состав формируемых метаданных должен обеспечивать возможность интеграции электронных коллекций в единое пространство научных знаний (см., например, [2, 3]). Важной составляющей метаданных является аффилиация авторов научных документов (см., например, [4]). В наукометрических системах аффилиация авторов публикаций используется для определения рейтинга научного журнала, а также как показатель

публикационной активности авторов статьи и научных учреждений, в которых они работают. Отсутствие какой-либо из частей аффилиации (названия организации, города, страны) ведет к ошибкам в индексировании, более того, статья с неполной аффилиацией, возможно, не будет проиндексирована в зарубежных базах данных. Кроме того, при индексации в наукометрических системах наличие полной аффилиации в статье позволяет правильно идентифицировать автора статьи, имеющего распространенную фамилию. Основные требования к составляющим аффилиации авторов научных публикаций, примеры влияния точности и полноты представленной в ней информации, приведены в [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 указана только фамилия с одним инициалом, что затрудняет подготовку метаданных аффилиации [6].

В работе представлен метод формирования метаданных, представляющих аффилиацию авторов публикаций и их нормализацию в форматах агрегирующих цифровых библиотек. Необходимые метаданные извлекаются из текстов статей сервисами фабрики метаданных цифровой библиотеки Lobachevskii DML и уточняются с помощью разработанных SPARQL-запросов из Wikidata и других открытых источников Сети.

2. Представление аффилиации авторов научного документа в метаданных документа

Для представления метаданных документов в цифровой математической библиотеке Lobachevskii DML используется формат, основанный на xml-схеме Journal Archiving and Interchange Tag Suite (NISO JATS) [7]. На рис. 1 приведен фрагмент метаданных документа, содержащий теги описания полной аффилиации автора документа.

```
<address>
    <institution-wrap>
        <institution content-type="uni"> организация </institution>
        <institution-id>идентификатор ROR</institution-id>
    </institution-wrap>

    <institution-wrap>
        <institution content-type="depe"> факультет </institution>
        <institution-id></institution-id>
    </institution-wrap>
    <addr-line> адрес </addr-line>
    <city> город </city>
    <postal-code> индекс </postal-code>
    <country> страна </country>
</address>
```

Рис. 1. Фрагмент метаописания афилиации автора документа в формате NISO JATS V1.3

Агрегирующие цифровые библиотеки предлагают собственные xml-схемы формирования метаданных (см., например, [8]). Эти схемы метаданных, в частности, содержат шаблоны для описания сведений об организации, в которой проведено исследование авторами статьи. Для преобразования метаданных в соответствии с этими схемами используются сервисы нормализации метаданных цифровой библиотеки Lobachevskii DML [9].

В работах [4, 5] указаны требования к составу и представлению аффилиации авторов научных документов. Также выделены 9 типов аффилиаций авторов, в соответствии с полнотой указанной информации об организации. Полная аффилиация авторов статьи включает в себя: название факультета, название организации, адрес (дом, улица), город, индекс, страна. На рис. 2 представлены примеры аффилиаций авторов документов, которые не являются полными по своему составу. В первом примере аффилиация является неполной, так как отсутствует указание страны. Во втором примере в аффилиации указано только научное учреждение. В третьем – организация и её подразделение. В четвертом примере документ не содержит аффилиации – указан только город.

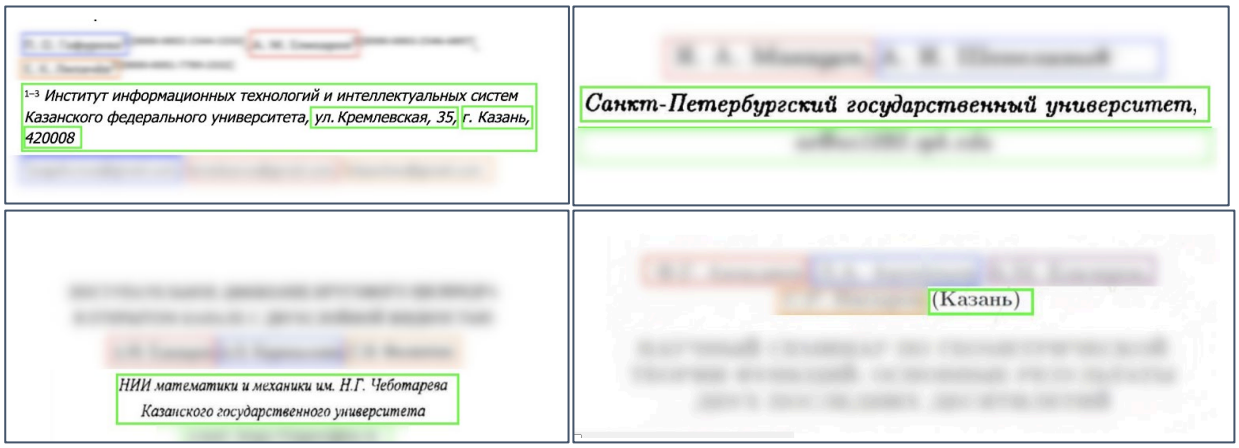

Рис. 2. Неполные аффилиаций авторов документов электронных коллекций

3. Алгоритм формирования метаданных аффилиации авторов научного документа

Приведем основные этапы алгоритма подготовки блока метаданных аффилиации авторов документа в соответствии с xml-схемами электронных коллекций цифровых математических библиотек.

На первом этапе производится поиск аффилиации авторов в тексте документа. Для этого используются программные инструменты, основанные на анализе структуры научного документа и алгоритмах поиска и извлечения основных блоков документа, в частности, аффилиации авторов [10, 11].

На следующем этапе производится определение типа аффилиации в соответствии с полнотой представленной в ней информации. Для этого в алгоритме используется такие как знаки препинания, цифры и капитализация символов.

Отдельным этапом алгоритма является определения языка, на котором представлен документ. Для этого используется посимвольный анализ текста.

Далее производится разделение аффилиации на составляющие: факультет или другое структурное подразделение, организация, дом, улица, город, почтовый индекс, страна. Для извлечения составляющих аффилиации разработаны шаблоны регулярных выражений, а также использованы специализированные библиотеки извлечения именованных сущностей [12, 13].

На следующем этапе выполняется анализ составляющих аффилиации на соответствие требованиям к подготовке метаданных. Прежде всего, осуществляется проверка наличия в аффилиации аббревиатур и сокращений. При обнаружении аббревиатур вызывается модуль замены на полное название организации. Для этого разработан соответствующий словарь, содержащий полное наименование организаций, принятые сокращенные наименования, а также идентификаторы ROR (Research Organization Registry, https://ror.org/). Для аффилиаций на английском языке вызывается функция, осуществляющая капитализацию начальных символов.

Далее приведен псевдокод алгоритма обработки аффилиации авторов научного документа.

Алгоритм 1. Формирование метаданных аффилиаций авторов в документе цифровой коллекции в формате NISO JATS V1.3

```
1:    ArPaper = Load() // загрузка файлов электронной коллекции
2:    for each Paper in ArPaper
3:      Affiliations[] = Find_Affiliation(Paper)
4:      Authors[] = Find_Authors(Paper)
5:      lang = DetectLang() // определение языка аффилиации
```

```
 6:      for each Aff in Affiliations
 7:        type = Affiliation_Type(Aff) // определение типа
аффилиации
 8:        Aff_Pattern = Pattern_Type(type) // подбор
шаблона регулярных выражений в соответствии с типом
аффилиации
 9:        Aff_Component[] = Split_Affiliation(Aff_Pattern)
// разбор аффилиации на структурные составляющие
 10:       if(Find_Abbreviation == True)
Correct_Abbreviation()
 11:       if(lang == “en”) Correct_Capitalization()
 12:       Meta_Affiliation(Aff_Component) // формирование
аффилиации в формате JATS
 13:     end for
 14:   end for
```

Для уточнения и пополнения информации о научных организациях используются открытые источники Семантической сети. Анализ представления информации о российских научных организациях в международных базах данных приведен в работе [14].

В таблице 1 приведены свойства Wikidata, использованные в алгоритме пополнения и уточнения метаданных аффилиации документов электронных коллекций.

Таблица 1. Свойства Wikidata, используемые в алгоритме пополнения аффилиации

Свойство	Wikidata EntityID	Описание	Тег JATS
official name	P1448	наименования организации (включая предыдущие названия);	<institution content-type= “uni”>
native label	P1705	наименования организации на языке страны расположения	<institution content-type= “uni” lang=“”>
short name	P1813	короткие наименования;	<institution content-type= “uni_short”>
country	P17	страна расположения организации	<country>

located in the administrative territorial entity	P131	территориальное расположение (город, адрес);	<city>, <addr-line>
headquarters location	P159	расположение главного здания (адрес)	<addr-line>
Replaces	P1365	ссылки на филиалы организации (в Wikidata)	
official website	P856	Официальный web сайт	<url>

На рис. 3 представлен фрагмент SPARQL-запроса уточнения составляющих аффилиации.

```
SELECT ?ofname ?short ?adress WHERE {
  ?item rdfs:label "Kazan Federal
  University"@en.
  ?item wdt:P31 wd:Q3918.
  ?item p:P1448 ?ofname        .
  ?item p:P1813 ?short  .
  ?item p:P159 ?adress.

}
```

Рис. 3. Извлечение информации из Wikidata для пополнения сведений об организации «Kazan Federal University»

Далее приведен псевдокод алгоритма пополнения составляющих аффилиации авторов документов электронных коллекций. Поиск необходимой информации осуществляется с помощью запросов в базу знаний Wikidata [15].

Алгоритм 2. Пополнение метаданных аффилиации авторов документов

```
1: Meta = Load_XML() //загрузить метаданные документа (полученные в Алгоритме 1)
2: Meta_Aff = XML_Path(Meta) // извлечение тегов аффилиации
3: Properties[] = Load_Properties() // Свойства из Табл. 1
4: Official_Name = XPath(<institution content-type= “uni”>
5: for each Tag in Meta_Aff
6:   Tag_Content = string(Tag)
7:    if Tag_Content == “” // отсутствует информация о
```

```
составляющей
8:     Result[][] = Query(Official_Name) //SPARQL-запрос к
Wikidata
9:     for each P in Properties //
10:      Meta_Aff=Add(Select(Result, P),Tag)
11:    end for
12:  end if
13: end for
14: Meta = Add(Meta_Aff)
15: Save(Meta)
```

На основе представленных алгоритмов реализован сервис, включенный в фабрику метаданных цифровой библиотеки Lobachevskii DML. Этот сервис был аппробирован на нескольких электронных коллекциях этой цифровой библиотеки.

Заключение

Предложен метод формирования метаданных аффилиаций авторов документов электронных коллекций, входящих в состав цифровой математической библиотеки Lobachevskii DML.

С помощью разработанной системы запросов к открытым источникам данных в алгоритме создаются семантические связи между составляющими аффилиации и сетевыми информационными объектами. Это позволило на основе последующего анализа извлеченной из сети информации, выполнить уточнение и пополнение метаданных аффилиации.

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

Литература

1. Elizarov A, Lipachev E., Big Math Methods in Lobachevskii-DML Digital Library // CEUR Workshop Proceedings. 2019. V. 2523. P. 59–72.
2. Атаева О.М., Каленов Н.Е., Серебряков В.А. Об основных понятиях Единого цифрового пространства научных знаний // Научный сервис в сети Интернет: труды XXII Всероссийской научной конференции (21–25 сентября 2020 г., онлайн). М.: ИПМ им. М.В.Келдыша, 2020. С. 29–40. https://doi.org/10.20948/abrau-2020-18
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 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
7. Journal Article Tag Suite. https://jats.nlm.nih.gov/about.html, last accessed 2020/12/12.
https://keldysh.ru/abrau/2021/theses/22.pdf, last accessed 2022/07/07.
8. EuDML metadata schema specification (v2.0–final). https://initiative.eudml.org/eudml-metadata-schema-specification-v20-final, last accessed 2022/07/07.
9. Гафурова П.О., Елизаров А.М., Липачев Е.К., Хамматова Д.М. Методы формирования и нормализации метаданных в цифровой математической библиотеке // Научный сервис в сети Интернет: труды XXI Всероссийской научной конференции (23-28 сентября 2019 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2019. С. 234–244. https://doi.org/10.20948/abrau-2019-28
10. Елизаров А.М., Липачёв Е.К., Хайдаров Ш.М. Автоматизированная система сервисов обработки больших коллекций научных документов // Аналитика и управление данными в областях с интенсивным использованием данных: сборник статей XVIII Междун. конф. DAMDID/RCDL'2016. М.: ФИЦ ИУ РАН, 2016. С.109–115.
11. Elizarov A.M., Khaydarov Sh.M., Lipachev E.K. Scientific documents ontologies for semantic representation of digital libraries // 2017 Second Russia and Pacific Conference on Computer Technology and Applications (RPC). Vladivostok, Russky Island, Russia 25-29 September, 2017. P. 1–5. https://doi.org/10.1109/RPC.2017.8168064
12. Lane H., Howard C., Hapke H. Natural Language Processing in Action. Manning Publications Co., 2019. 545 p.
13. Кукушкин А. Проект Natasha. Набор качественных открытых инструментов для обработки естественного русского языка (NLP), https://habr.com/ru/post/516098/

14. Апанович З.В. Сопоставление и интеграция информации о российских научных организациях из разноязычных источников данных // Научный сервис в сети Интернет: труды XXIII Всероссийской научной конференции (20-23 сентября 2021 г., онлайн). — М.: ИПМ им. М.В.Келдыша, 2021. — С. 34-42. https://doi.org/10.20948/abrau-2021-13
15. Wikidata. URL: https://www.wikidata.org/wiki/Wikidata:Main_Page

Reference

1. Elizarov A, Lipachev E., Big Math Methods in Lobachevskii-DML Digital Library // CEUR Workshop Proceedings. 2019. V. 2523. P. 59–72.
2. Ataeva O.M., Kalenov N.E., Serebryakov V.A. The basic concepts of a common digital space of scientific knowledge // Nauchnyj servis v seti Internet: trudy XXII Vserossijskoj nauchnoj konferencii (21–25 sentyabrya 2020 g., online). M.: IPM im. M.V.Keldysha, 2020. S. 29–40. https://doi.org/10.20948/abrau-2020-18
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. Significance and Basic Affiliation Requirements in Scientific Publications // Science Editor and Publisher. 2016. V. 1 (1–4). P. 32–42.
6. 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 // Nauchnyj servis v seti Internet: trudy XXIII Vserossijskoj nauchnoj konferencii (20–23 sentyabrya 2021 g., online). M.: IPM im. M.V.Keldysha, 2021. S. 22–33. https://doi.org/10.20948/abrau-2021-22
7. Journal Article Tag Suite. https://jats.nlm.nih.gov/about.html, last accessed 2020/12/12. https://keldysh.ru/abrau/2021/theses/22.pdf, last accessed 2022/07/07.
8. EuDML metadata schema specification (v2.0–final). https://initiative.eudml.org/eudml-metadata-schema-specification-v20-final, last accessed 2022/07/07.
9. Gafurova P.O., Elizarov A.M., Lipachev E.K., Khammatova D.M. Methods of Formation and Normalization of Metadata in the Digital Mathematical Library // Nauchnyj servis v seti Internet: trudy XXI Vserossijskoj nauchnoj

konferencii (23–28 sentyabrya 2019 g., g. Novorossiisk). sentyabrya, 2019. S. 234–244. https://doi.org/10.20948/abrau-2019-28
10. Elizarov A.M., Lipachev E.K., Khaydarov Sh.M. Avtomatizirovannaya sistema servisov obrabotki bol'shih kollekcij nauchnyh dokumentov // Analitika i upravlenie dannymi v oblastyah s intensivnym ispol'zovaniem dannyh: sbornik statej XVIII Mezhdun. konf. DAMDID/RCDL'2016. M.: FIC IU RAN, 2016. C.109–115.
11. Elizarov A.M., Khaydarov Sh.M., Lipachev E.K. Scientific documents ontologies for semantic representation of digital libraries // 2017 Second Russia and Pacific Conference on Computer Technology and Applications (RPC). Vladivostok, Russky Island, Russia 25-29 September, 2017. P. 1–5. https://doi.org/10.1109/RPC.2017.8168064
12. Lane H., Howard C., Hapke H. Natural Language Processing in Action. Manning Publications Co., 2019. 545 p.
13. Kukushkin A. Proekt Natasha. Nabor kachestvennyh otkrytyh instrumentov dlya obrabotki estestvennogo russkogo yazyka (NLP), https://habr.com/ru/post/516098/
14. Apanovich Z.V. Matching and integration of data about Russian research organizations from multilingual data sources // CEUR Workshop Proceedings. 2021. V. 3066. P. 133-139.
15. Wikidata. URL: https://www.wikidata.org/wiki/Wikidata:Main_Page