Структура и сервисы фабрики метаданных цифровой математической библиотеки

П.О. Гафурова[1][0000-0002-1544-155X], А.М. Елизаров[1,2][0000-0003-2546-6897], Е.К. Липачёв[1,2][0000-0001-7789-2332]

[1]Институт математики и механики им. Н.И. Лобачевского
[2]Высшая школа информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета

Аннотация. Представлены решения ряда задач управления метаданными, возникающих при построении цифровых математических библиотек. Описаны структура фабрики метаданных цифровой математической библиотеки Lobachevskii-DML и набор сервисов, составляющих основу этой фабрики. Фабрика метаданных, в данной статье, определяется как система взаимосвязанных программных инструментов, направленных на создание, обработку, хранение и управление метаданными объектов цифровых библиотек и позволяющих интегрировать создаваемые электронные коллекции в агрегирующие цифровые научные библиотеки.

Ключевые слова: цифровая библиотека, цифровая математическая библиотека, фабрика метаданных, сервисы управления метаданными, цифровая математическая библиотека Lobachevskii-DML. Structure and Services of the Digital Mathematical Library

Structure and Services of the Digital Mathematical Library Metadata Factory

P.O. Gafurova[1][0000-0002-1544-155X], A.M. Elizarov[1,2][0000-0003-2546-6897], E.K. Lipachev[1,2][0000-0001-7789-2332]

[1] N. I. Lobachevskii Institute of Mathematics and Mechanics,
[2]Higher School of Information Technologies and Intelligent Systems, Kazan (Volga Region) Federal University

Abstract. The solutions of a number of metadata management problems arising in the construction of digital mathematical libraries are presented. The structure of the metadata factory of the digital mathematical library Lobachevskii-DML and the set of services that form the basis of this factory are described. The metadata factory, in this

article, is defined as a system of interconnected software tools aimed at creating, processing, storing and managing metadata of digital library objects and allowing integrating created electronic collections into aggregating digital scientific libraries.

Keywords: digital library, digital mathematical library, metadata factory, metadata management services, digital mathematical library Lobachevskii-DML.

1. Введение

Как хорошо известно (см., например, [1]), важной составляющей информационного научного пространства являются цифровые библиотеки. Решение основных задач, поставленных в проектах интеграции математических знаний, связывают с развитием цифровых математических библиотек [2–4]. Обзор специфики и функциональных возможностей ряда существующих цифровых математических библиотек содержится в [5].

В настоящей статье представлены базовые сервисы фабрики метаданных цифровой математической библиотеки Lobachevskii-DML. Для решения задачи интеграции информационных ресурсов предложены методы преобразования мета-данных электронных коллекций и содержащихся в них документов по DTD-правилам и XML-схемам Journal Archiving and Interchange Tag Suite (NISO JATS) различных версий [6].

Мы используем термин фабрика метаданных цифровой библиотеки (metadata factory of digital library) в том же смысле, в каком он использован в [7], а именно: фабрика метаданных – это система взаимосвязанных программных инструментов, направленных на создание, обработку, хранение и управление метаданными объектов цифровых библиотек и позволяющих интегрировать создаваемые электронные коллекции в агрегирующие цифровые научные библиотеки. С помощью этих инструментов преимущественно в автоматизированном режиме выполняются такие операции, как выделение объектов и связей между ними, экстракция метаданных из различных источников и конкретных документов, верификация, уточнение, улучшение, нормализация в различных форматах и гармонизация метаданных с помощью ручного редактирования или автоматизированных агентов, хранение и связывание метаданных с внешними базами данных. В случае цифровой математической библиотеки к перечисленным инструментам добавляется ряд специализированных, таких, например, как преобразование в формат MathML, разметка математических формул и организация поиска по ним [8].

2. Сервисы управления цифровым математическим контентом

Создание цифровой математической библиотеки и последующее расширение её функциональных возможностей предполагают решение целого ряда трудоемких задач, связанных, в том числе, с управлением контентом (см., например,

[1]). Структура наиболее известных цифровых математических библиотек и разработанные в них сервисы управления документами и коллекциями обсуждены в [4, 5].

Важной составляющей любой цифровой библиотеки являются программные инструменты управления научным контентом. Эти инструменты используются и фабрикой метаданных для создания, обработки, хранения и управления метаданными электронных документов и позволяют интегрировать создаваемые электронные коллекции в агрегирующие цифровые научные библиотеки. Опишем подробнее существующие решения.

Существующие цифровые библиотеки, а также агрегаторы научных знаний предлагают ряд программных инструментов работы с контентом, прежде всего, сервисы поиска в электронных коллекциях. Например, средства семантического поиска документов представлены на сайте проекта EuDML (https://initiative. eudml.org/). Здесь же размещены демонстрационные версии инструментов, разработанных для обслуживания EuDML. Назначение и функциональные возможности этих программных инструментов описаны в [9].

Сервисы, разработанные и реализованные в рамках проекта Lobachevskii-DML описаны в работах [10–12].

3. Фабрика метаданных цифровой математической библиотеки

В настоящем разделе мы предлагаем решения ряда задач, связанных с построением фабрики метаданных в рамках проекта создания цифровой математической библиотеки Lobachevskii-DML [13]. Как и в случае любой цифровой научной библиотеки, формирование библиотеки Lobachevskii-DML и соответствующей фабрики метаданных потребовало привлечения ранее созданных, а также разработки новых технологических решений управления научным контентом.

На этапе препроцессорной обработки выполняется отсев тех документов, которые не получилось обработать в автоматическом режиме, с указанием возникших проблем в файле отчета, сформированном автоматически. Также на этом этапе производятся исправление некоторых ошибок орфографии, неправильного выбора регистра, а также удаление лишних пробелов и знаков.

На этапе экстракции метаданных обрабатываются полные тексты документов, используются шаблоны поиска обязательных метаданных.

На этапе верификации метаданных выполняется проверка полноты и соответствия состава выделенных метаданных установленным правилам, записанным в виде DTD-файлов или XML-схем. После прохождения этапа верификации возможны три варианта дальнейших действий: дополнительная экстракция необходимых и дополнительных метаданных; повторная верификация метаданных и выдача отчета о том, что документ недостаточен для получения требуемых метаданных; переход к финальному действию – нормализации метаданных.

Экстракция дополнительных метаданных направлена на извлечение метаданных не только из самого документа, но и с помощью внешних ресурсов (например, персональных страниц авторов).

В следующем разделе представлены некоторые из уже реализованных нами инструментов фабрики метаданных.

4. Сервисы экстракции метаданных в фабрике метаданных цифровой математической библиотеки

Размещение метаданных в интернете привело к тому, что одним из их источников могут стать веб-страницы сайта-агрегатора метаданных или самой цифровой библиотеки. Таким образом, при формировании фундаментального набора метаданных электронных коллекций, а также при получении дополнительных метаданных необходимо использовать метаданные, хранящиеся на внешних ресурсах. Эта задача сопряжена с задачами поиска информации в агрегирующих базах данных и цифровых библиотеках, некоторые из которых частично закрыты для доступа или прерывают соединение, позволяя скачивать только ограниченное количество метаданных. При поиске метаданных на страницах сайтов-агрегаторов нужно также понимать и учитывать, что выбор и порядок поиска в таких источниках должны быть определены заранее, так как некоторые источники хранят информацию только по конкретной тематике (например, библиографическая база данных DBLP) или же неполный список метаданных.

Один из случаев экстракции метаданных с сайтов-агрегаторов разработан нами на основе сайта проекта MathNet. Цель созданного программного приложения – выделение и запись метаданных статьи на русском и английском языках с дальнейшей нормализацией по формату, принятому в EuDML. Основные шаги алгоритма экстракции и нормализации метаданных на примере одной из коллекций приведены в [12].

5. Сервисы нормализации метаданных в фабрике метаданных цифровой математической библиотеки

Задача перевода метаданных из одного формата в другой связана с задачами дополнения и улучшения метаданных. В цифровой библиотеке Lobachevskii DML реализован сервис перевода метаданных электронной коллекции статей журнала «Электронные библиотеки» (“Russian Digital Libraries Journal”, https://elbib.ru/) в формат базы данных DBLP. Процесс перевода включал семантическую транслитерацию имен и фамилий авторов статей. Исходные наборы метаданных, использованных при переводе в названной формат, были сформированы автоматически с помощью разработанных нами программных инструментов, с учётом специфики программной платформы OJS [14], на которой функционирует данный журнал. Алгоритм перевода этих метаданных в формат DBLP был успешно реализован, подробно он подробно в [11, 12].

Заключение

Описана структура фабрики метаданных цифровой математической библиотеки Lobachevskii-DML. Представлен набор сервисов, составляющих основу этой фабрики. Фабрика метаданных определена как система взаимосвязанных программных инструментов, направленных на создание, обработку, хранение и управление метаданными объектов цифровых библиотек и позволяющих интегрировать создаваемые электронные коллекции в агрегирующие цифровые научные библиотеки.

Дальнейшее направление развития заключается в совершенствовании созданной фабрики метаданных и разработке возможности ее использования в любых научных цифровых библиотеках.

Работа выполнена при частичной финансовой поддержке РФФИ и Правительства Республики Татарстан в рамках научного проекта № 18-47-160012. Настоящая статья содержит также результаты проекта «Разработка технологий управления математическими знаниями на основе цифровой математической библиотеки Lobachevskii-DML», выполняемого в рамках реализации Программы Центра компетенций Национальной технологической инициативы «Центр хранения и анализа больших данных», поддерживаемого Министерством науки и высшего образования Российской Федерации по Договору МГУ им. М.В. Ломоносова с Фондом поддержки проектов Национальной технологической инициативы от 11.12.2018 № 13/1251/2018.

Литература

1. Xie I., Matusiak K.K. Discover Digital Libraries: Theory and Practice. Elsevier Inc., 2016. – 388 p.
2. Jackson A. The Digital Mathematics Library // Notices Amer. Math. Soc. –2003. – V. 50. – P. 918–923.
3. Borwein J.M., Rocha E.M., Rodrigues J.F. (eds.) Communicating Mathematics in the Digital Era. Taylor & Francis, 2008. – 325 p.
4. Bouche T. The Digital Mathematics Library as of 2014 // Notices Amer. Math. Soc. – 2014. – V. 61. – No 9. – P. 1085–1088.
5. Elizarov A.M., Lipachev E.K., Zuev D.S. Digital Mathematical Libraries: Overview of Implementations and Content Management Services // CEUR Workshop Proceedings. – 2017. – V. 2022. – P. 317–325.
6. Journal Article Tag Suite. URL: https://jats.nlm.nih.gov/about.html.
7. Bouche T., Labbe O. The New Numdam Platform // In: Geuvers H., England M., Hasan O., Rabe F., Teschke O. (Eds) Intelligent Computer Mathematics. CICM 2017. Lecture Notes in Computer Science. V. 10383. Springer, Cham, 2017. – P. 70–82. – URL: https://doi.org/10.1007/978-3-319-62075-6_6. https://zenodo.org/ record/581405/.

8. Елизаров А.М., Липачёв Е.К., Невзорова О.А., Соловьев В.Д. Методы и средства семантического структурирования электронных математических документов // Доклады РАН. – 2014. – Т. 457, № 6.– С. 642–645. – URL: https://doi.org/ 10.7868/ S0869565214240049.
9. D7.4: Toolset for Image and Text Processing and Metadata Enhancements — Final Release. URL: https://wiki.eudml.eu/mediawiki/eudml/ images/D7.4-v1.0.pdf.
10. Елизаров А.М., Липачёв Е.К. Семантические методы и инструменты электронной математической библиотеки Lobachevskii-DML // Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (18–23 сентября 2017 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2017. – С. 130–136. – URL: https://doi.org/10.20948/abrau-2017-73. http://keldysh.ru/ abrau/ 2017/73.pdf.
11. .Гафурова П.О., Елизаров А.М., Липачёв 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. URL: https://doi.org/10.20948/abrau-2019-28. http://keldysh.ru/abrau/2019/theses/28.pdf.
12. Gafurova P.O., Elizarov A.M., Lipachev E.K., Khammatova D.M. Metadata Normalization Methods in the Digital Mathematical Library // CEUR Workshop Proceedings. – 2020. – V. 2543. – P. 136–148.
13. Elizarov A.M., Lipachev E.K. Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // CEUR Workshop Proceedings. – 2017. – V. 2022. – P. 326–333.
14. MacGregor J., Stranack K., Willinsky J. The Public Knowledge Project: Open Source Tools for Open Access to Scholarly Communication // In: Bartling S., Friesike S. (Eds) Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer International Publishing, 2014. – P. 165–175. – URL: https://doi.org/10.1007/978-3-319-00026-8_3.

Reference

1. Xie I., Matusiak K.K. Discover Digital Libraries: Theory and Practice. Elsevier Inc., 2016. – 388 p.
2. Jackson A. The Digital Mathematics Library // Notices Amer. Math. Soc. –2003. – V. 50. – P. 918–923.
3. Borwein J.M., Rocha E.M., Rodrigues J.F. (eds.) Communicating Mathematics in the Digital Era. Taylor & Francis, 2008. – 325 p.

4. Bouche T. The Digital Mathematics Library as of 2014 // Notices Amer. Math. Soc. – 2014. – V. 61. – No 9. – P. 1085–1088.
5. Elizarov A.M., Lipachev E.K., Zuev D.S. Digital Mathematical Libraries: Overview of Implementations and Content Management Services // CEUR Workshop Proceedings. – 2017. – V. 2022. – P. 317–325.
6. Journal Article Tag Suite. URL: https://jats.nlm.nih.gov/about.html.
7. Bouche T., Labbe O. The New Numdam Platform // In: Geuvers H., England M., Hasan O., Rabe F., Teschke O. (Eds) Intelligent Computer Mathematics. CICM 2017. Lecture Notes in Computer Science. V. 10383. Springer, Cham, 2017. – P. 70–82. – URL: https://doi.org/10.1007/978-3-319-62075-6_6. https://zenodo.org/ record/581405/.
8. Elizarov A.M., Lipachev E.K., Nevzorova O.A., Solov'ev V.D. Methods and Means for Semantic Structuring of Electronic Mathematical Documents // Doklady Mathematics. 2014. 90 (1). P. 521–524. URL: https://doi.org/0.1134/ S1064562414050275.
9. D7.4: Toolset for Image and Text Processing and Metadata Enhancements — Final Release. URL: https://wiki.eudml.eu/mediawiki/eudml/ images/D7.4-v1.0.pdf.
10. Elizarov A.M., Lipachev E.K. Semanticheskie metody` i instrumenty` e`lektronnoj matematicheskoj biblioteki Lobachevskii-DML // Nauchny`j servis v seti Internet: trudy` XIX Vserossijskoj nauchnoj konferencii (18–23 sentyabrya 2017 g., g. Novorossijsk). M.: IPM im. M.V. Keldy`sha, 2017. S. 130–136. https://doi.org/ 10.20948/abrau-2017-73. URL: http://keldysh.ru/abrau/2017/ 73.pdf.
11. 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. 2019. – S. 234–244. https://doi.org/10.20948/abrau-2019-28. URL: http://keldysh.ru/abrau/ 2019/theses/28.pdf.
12. Gafurova P.O., Elizarov A.M., Lipachev E.K., Khammatova D.M. Metadata Normalization Methods in the Digital Mathematical Library // CEUR Workshop Proceedings. – 2020. – V. 2543. – P. 136–148.
13. Elizarov A.M., Lipachev E.K. Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // CEUR Workshop Proceedings. – 2017. – V. 2022. – P. 326–333.
14. MacGregor J., Stranack K., Willinsky J. The Public Knowledge Project: Open Source Tools for Open Access to Scholarly Communication // In: Bartling S., Friesike S. (Eds) Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer International Publishing, 2014. – P. 165–175. – URL: https://doi.org/10.1007/978-3-319-00026-8_3.