**УДК 004.4**
**ББК 30ф**

**Гафурова П.О.[1], Елизаров А.М.[2], Липачёв Е.К.[3]**
Институт математики и механики им. Н.И. Лобачевского
Высшая школа информационных технологий
и интеллектуальных систем
Казанского (Приволжского) федерального университета
Казань, Россия
[1] polina.mannshtern@mail.ru, [2] amelizarov@gmail.com,
[3] elipachev@gmail.com

# МЕТОДЫ НОРМАЛИЗАЦИИ МЕТАДАННЫХ ЭЛЕКТРОННЫХ МАТЕМАТИЧЕСКИХ КОЛЛЕКЦИЙ

***Аннотация**: Предложены методы интеграции электронных математических коллекций Казанского университета в цифровые математические библиотеки. Представлены алгоритмы пополнения электронных коллекций цифровой математической библиотеки Lobachevskii-DML и формирования метаданных в форматах этой библиотеки. Разработаны сервисы нормализации метаданных коллекций этой цифровой библиотеки в соответствии с XML-схемами NISO JATS и DBLP. Представлены алгоритмы создания обязательного и фундаментального наборов метаданных коллекций указанной библиотеки в соответствии с правилами европейской цифровой математической библиотеки EuDML.*

***Ключевые слова**: электронные библиотеки, извлечение метаданных, семантические связи информационных объектов, цифровая математическая библиотека Lobachevskii-DML.*

**Gafurova P.O.[1], Elizarov A.M.[2], Lipachev E.K.[3]**
N.I. Lobachevski Institute of Mathematics and Mechanics
Higher School of Information Technologies
and Intelligent Systems
Kazan (Volga region) Federal University
Kazan, Russia
[1] polina.mannshtern@mail.ru, [2] amelizarov@gmail.com,
[3] elipachev@gmail.com

# METHODS OF NORMALIZATION OF DIGITAL MATHEMATICAL COLLECTIONS METADATA

***Abstract**: We offer methods for integrating electronic mathematical collections of the Kazan University into digital mathematical libraries. We provide algorithms for the replenishment of the electronic collections of the digital mathematical library Lobachevskii-DML, as well as methods for generating metadata in the formats of this library. We offer metadata services for the collections of this digital library in accordance with the NISO JATS and DBLP XML schemas. We describe the algorithms for creating the mandatory and fundamental metadata sets of the collections of the specified library according to the requirements of the EuDML digital mathematical library.*

***Keywords**: digital libraries, extraction of metadata, semantic relations of information objects, Lobachevskii Digital Mathematical Library.*

## Введение

Система цифровых математических библиотек, создаваемая в настоящее время, призвана консолидировать и сделать доступными как современные математические знания, так и документы, опубликованные ранее. Для достижения этой цели в рамках цифровых библиотек разрабатываются методы управления цифровой информацией, учитывающие особенности представления математического контента (см., напр., [1–4]). Одной из таких библиотек является Lobachevskii Digital Mathematical Library (Lobachevskii-DML, https://lobachevskii-dml.ru/), создаваемая в Казанском университете [5–7]. Особенностью этой библиотеки является использование семантических методов обработки математического контента, основанных на управлении объектами и связями между ними [8, 9]. Ещё одно направление развития этой библиотеки связано с созданием методов автоматизированной обработки больших массивов документов [10, 11].

В настоящей статье представлены методы формирования метаданных электронных коллекций библиотеки Lobachevskii-DML. Предложены методы преобразования метаданных для интеграции создаваемых электронных коллекций в глобальное информационное

математическое пространство в рамках инициатив World Digital Mathematical Library (WDML) [8], The Global Digital Mathematics Library (GDML) [12] и проектов The European Digital Mathematics Library (EuDML, https://initiative.eudml.org/) [13, 14], MathNet.Ru (http://www.mathnet.ru/) [1]. Описаны сервисы нормализации метаданных коллекций цифровой библиотеки Lobachevskii-DML в соответствии с DTD-правилами и XML-схемами Journal Archiving and Interchange Tag Suite (NISO JATS V1.0, V1.1, V1.2, https://jats.nlm.nih.gov/archiving/) [15]. Представлены алгоритмы создания обязательного и фундаментального наборов метаданных в соответствии с правилами европейского интегратора цифровых математических знаний EuDML. Приведен также алгоритм автоматизированной подготовки метаданных электронных коллекций библиотеки Lobachevskii-DML по правилам библиографической базы по компьютерным наукам "Dblp Computer Science Bibliography" (DBLP, https://dblp.uni-trier.de/).

**Форматы метаданных цифровых математических библиотек и методы нормализации метаданных**

В настоящее время публикации по математике индексируются во многих наукометрических базах данных (см., напр., [16]). Эти базы предъявляют различные требования к типу документов (например, как правило, не индексируются новые формы публикаций: презентации, научные блоги, видеолекции), составу метаданных этих документов и схемам их представления.

Цифровые математические библиотеки при формировании коллекций, входящих в них, используют несколько форматов метаданных, что отчасти связано с тем, что в состав этих коллекций включаются документы, созданные по правилам, использованным ранее и уже устаревшим, например, архивы научных журналов. Поэтому возникает необходимость в разработке методов извлечения из документов метаданных, а также методов приведения уже созданных метаданных в форматы соответствующих наукометрических баз данных. Участие в интегрирующих проектах, таких, как EuDML, предполагает предоставление наборов метаданных, сформированных по схемам агрегаторов математических ресурсов.

Термин *нормализация* используется нами для обозначения методов формирования или преобразования метаданных документов в соответствии с правилами и XML-схемами цифровых библиотек и наукометрических баз данных.

Схема метаданных цифровой математической библиотеки EuDML описана в [17]. Метаданные разделены на основные, фундаментальные и дополнительные [18]. Для описания журнальных статей

в проекте EuDML используются XML-схемы (NISO JATS V1.0) [19]. Обязательный набор метаданных EuDML является минимальным по составу и содержит название статьи на языке оригинала, фамилии и имена авторов, список библиографии, уникальный идентификатор статьи, например, doi, URL полного текста статьи. Фундаментальный набор метаданных дополнительно к обязательным метаданным включает аннотацию статьи и ключевые слова.

Ряд электронных коллекций библиотеки Lobachevskii-DML физически размещен в других цифровых библиотеках. Например, журнальная коллекция «Известия вузов. Математика» («Russian Mathematics (Izvestiya VUZ. Matematika)») оцифрована, снабжена метаописаниями и представлена на портале MathNet.Ru (http://www.mathnet.ru/php/journal.phtml?jrnid=ivm). Нашими задачами являются пополнение таких коллекций дополнительными метаданными, а также автоматическое выделение объектов и установление семантических связей между ними.

При формировании фундаментального набора метаданных электронных коллекций, хранящихся на внешних ресурсах, первоначально производится импорт метаданных, представленных на этих ресурсах. Для этой цели на языке C# и с использованием функций пакета HtmlAgilityPack (https://html-agility-pack.net/) разработана программа выделения метаданных с веб-страниц и их записи в XML-формате цифровой библиотеки Lobachevskii-DML, их пополнения и последующего преобразования по схемам EuDML. Например, для коллекции «Известия вузов. Математика» выполнены следующие шаги. Часть метаданных импортирована из блока «Цитирование в формате AMSBIB», представленного на веб-странице статьи портала MathNet.Ru. Затем с веб-страницы считаны ключевые слова и гиперссылка на страницу портала Springer Link (https://link.springer.com/ journal/11982) с англоязычной версией статьи. Эта информация включена в состав метаданных, и выполнен переход по гиперссылке. Следующий шаг состоял из анализа веб-страницы англоязычной версии статьи, извлечения и записи метаданных. Далее был генерирован персональный идентификатор этой статьи, который предложено создавать как объединение строк — идентификатора журнала (значение атрибута «`jrnid=`») и идентификатора статьи (значение атрибута «`paperid=`») на портале MathNet.Ru.

В настоящее время схемы, предложенные EuDML, не позволяют соединить в рамках одного метаописания статью, опубликованную на русском языке, и ее переводную версию на английском языке. Например, статья «Авхадиев Ф.Г., Насибуллин Р.Г., Шафигуллин И.К. // Известия высших учебных заведений. Математика. 2018.

№8. С. 88–92» и ее перевод на английский язык “Avkhadiev F.G., Nasibullin R.G., Shafigullin I.K. // Russian Mathematics. 2018. V. 62. No 8. P. 76–79” в фундаментальном наборе EuDML приходится описывать как различные статьи в разных журналах. В коллекциях Lobachevskii-DML, а также библиотеках eLibrary.ru и MathNet.ru такие статьи представлены как дубликаты одного документа.

В настоящий момент наукометрической базой, осуществляющей индексацию публикаций исключительно по компьютерной тематике, является “Dblp Computer Science Bibliography” (DBLP, https://dblp.uni-trier.de/) [16]. Среди коллекций цифровой библиотеки Lobachevskii-DML наиболее подходящей требованиям DBLP по тематике является коллекция журнала «Электронные библиотеки» (“Russian Digital Libraries Journal”, https://elbib.ru/). Включение этой коллекции в наукометрическую базу DBLP рассматривалось как необходимый шаг в развитии журнала.

В 2015 году редакцией журнала «Электронные библиотеки» была выбрана новая модель представления документов, существенно расширен состав метаданных и внедрена издательская система Open Journal System (OJS) [20]. Прием статей и последующие издательские процессы в этом журнале выполняются сегодня через систему OJS, а наборы метаданных формируются автоматически с помощью разработанных в редакции программных инструментов (http://ojs.kpfu.ru/index.php/elbib). Поэтому архив статей, вышедших в период с 2015 по 2018 годы, был выбран для подготовки к индексации в указанной базе данных. По требованиям этой наукометрической базы данных метаописание документа включает: идентификатор публикации, фамилии и имена авторов, название работы, год издания, том, номер, начальную и конечную страницы статьи в номере журнала, URL полного текста статьи. Метаданные документа представляются в форматах XML-dblp, BibTeX, Research Information Systems (RIS), RDF N-Triples и RDF/XML.

Формирование метаданных по схемам DBLP проводится в три этапа: экстракция требуемых метаданных, дополнение метаданных и их нормализация в указанные форматы. С помощью программы, разработанной на языке C# и средств расширения System:XML, выполняется последовательная обработка файлов коллекции и, как результат, формируется набор метаданных каждого документа. На следующем этапе метаданные пополняются информацией о статье и её авторах на английском языке. Эта информация импортируется с англоязычной версии сайта журнала. Поскольку англоязычная информация об авторах неполная — указаны только фамилии и инициалы — производится транслитерация имен с русскоязычной страницы.

Результатом этой работы стало включение журнала “Russian Digital Libraries Journal” и статей, вышедших в 2015–2018 годах, в базу DBLP (https://dblp.uni-trier.de/db/journals/rdlj/).

**Заключение**

С целью интеграции электронных математических коллекций Казанского университета в международное научное пространство разработаны алгоритмы формирования метаданных этих коллекций и документов, входящих в них, в соответствии с форматами цифровых математических библиотек и наукометрических баз данных. Представлены методы нормализации метаданных электронных математических коллекций в соответствии с XML-схемами NISO JATS и DBLP.

**Благодарность**

Настоящая статья содержит 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.

**Источники:**

[1] Chebukov D.E., Izaak A.D., Misyurina O.G., Pupyrev Yu.A., and Zhizhchenko A.B. Math-Net.Ru as a Digital Archive of the Russian Mathematical Knowledge from the XIX Century to Today. Intelligent Computer Mathematics. // LNCS. 2013. Vol. 7961. P. 344–348. https://doi.org/10.1007/978-3-642-39320-4_26.

[2] Bartošek M., Rákosník J. DML-CZ: The Experience of a Medium-Sized Digital Mathematics Library // Notices of the AMS. 2013. Vol. 60. No 8. P. 1028–1033. http://dx.doi.org/10.1090/noti1031.

[3] Elizarov A.M., Lipachev E.K., Zuev D.S. Digital Mathematical Libraries: Overview of Implementations and Content Management Services // CEUR Workshop Proceedings. 2017. Vol. 2022. P. 317–325.

[4] Bouche T., Labbe O. The New Numdam Platform // CICM 2017: Intelligent Computer Mathematics, 2017. P. 70–82. https://doi.org/10.1007/978-3-319-62075-6_6.

[5] Elizarov A.M., Lipachev E.K. Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // CEUR Workshop Proceedings. 2017. Vol. 2022. P. 326–333.

[6] Елизаров А.М., Липачёв Е.К., Хайдаров Ш.М. Структура и сервисы цифровой математической библиотеки Lobachevskii-DML // Ученые записки ИСГЗ. 2017. Т. 15. № 1. С. 215–220.
[7] 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. Москва: ИПМ им. М. В. Келдыша, 2017. С. 130–136. https://doi.org/10.20948/abrau-2017-73.
[8] Developing a 21st Century Global Library for Mathematics Research. Washington: The National Academies Press, 2014. 131 p.
[9] Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A. Mathematical Knowledge Management: Ontological Models and Digital Technology // CEUR Workshop Proceedings. 2016. Vol. 1752. P. 44–50.
[10] Elizarov A.M., Lipachev E.K., and Khaidarov Sh.M. Automated Processing Service System of Large Collections of Scientific Documents // CEUR Workshop Proceedings. 2016. Vol. 1752. P. 58–64.
[11] Elizarov A.M., Khaydarov Sh.M., Lipachev E.K. Scientific Documents Ontologies for Semantic Representation of Digital Libraries // Proc. of the 2nd Russia and Pacific Conf. on Computer Technology and Applications. 2017. P. 1–5. https://doi.org/10.1109/RPC.2017.8168064.
[12] Ion P.D.F., Watt S.M. The Global Digital Mathematics Library and the International Mathematical Knowledge Trust // ICM 2017: Intelligent Computer Mathematics, 2017. Lecture Notes in Artificial Intelligence. Springer. 2017. Vol. 10383. P. 56–69. https://doi.org/10.1007/978-3-319-62075-6_5.
[13] Bouche T. Reviving the free public scientific library in the digital age? [Электр. ресурс]. The EuDML project // Kaiser K., Krantz S.G., Wegner B. (Eds.) Topics and Issues in Electronic Publishing JMM/AMS Special Session. FIZ Karlsruhe, 2013. P. 57–80. URL: https://www.emis.de/proceedings/TIEP2013/05bouche.pdf.
[14] Bouche T., Rákosník J. Report on the EuDML External Cooperation Model. [Электр. ресурс]. // Kaiser K., Krantz S.G., Wegner B. (Eds.) Topics and Issues in Electronic Publishing, JMM, Special Session, San Diego, 2013. P. 99–108. URL: https://www.emis.de/proceedings/TIEP2013/07bouche_rakosnik.pdf.
[15] "ANSI/NISO Z39.96-2019, JATS: Journal Article Tag Suite". [Электр. ресурс]. National Information Standards Organization. 8 February 2019. 652 p. URL: https://groups.niso.org/apps/group_public/download.php/21030/ANSI-NISO-Z39.96-2019.pdf.
[16] 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 [Электр. ресурс] // 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. URL: http://keldysh.ru/abrau/2018/theses/53.pdf doi:10.20948/abrau-2018-53.
[17] Jost M., Bouche T., Goutorbe C., Jorda J.P. D3.2: The EuDML metadata schema. [Электр. ресурс]. URL: http://www.mathdoc.fr/publis/d3.2-v1.6.pdf.

[18] EuDML metadata schema specification (v2.0–final). [Электр. ресурс]. URL: https://initiative.eudml.org/eudml-metadata-schema-specification-v20-final.
[19] Journal Article Tag Suite. [Электр. ресурс]. NISO JATS V1.0. URL: https://jats.nlm.nih.gov/1.0/.
[20] Ахметов Д.Ю., Елизаров А.М., Липачёв Е.К. Сервис-ориентированная информационная система научного журнала «Электронные библиотеки» // Электронные библиотеки. 2016. Т. 19. № 1. С. 2–39.