УДК 004.4
ББК 3

Елизаров А.М., Липачёв Е.К., Хайдаров Ш.М.
Казанский (Приволжский) федеральный университет
Казань, Россия
mailto:amelizarov@gmail.com, elipachev@gmail.com, 15jkeee@gmail.com

АВТОМАТИЗИРОВАННАЯ СИСТЕМА СТРУКТУРНОЙ И СЕМАНТИЧЕСКОЙ ОБРАБОТКИ ФИЗИКО-МАТЕМАТИЧЕСКОГО КОНТЕНТА

Аннотация: Представлена система сервисов автоматической обработки больших коллекций физико-математических документов. Сервисы обеспечивают: автоматическую валидацию документов и их преобразование в соответствии с правилами формирования коллекций (в частности, правилами представления статей в научные журналы); семантический анализ документов, извлечение метаданных, а также подготовку различных типов изданий научных материалов с выбором и дальнейшей корректировкой их структуры. Названные сервисы позволяют автоматически выполнять при обработке больших коллекций электронных документов такой набор операций и действий, который не реализуем при традиционной «ручной» работе с электронным контентом. Приведен один из примеров успешного использования системы.

Ключевые слова: *извлечение информации, интеграция данных, системы управления информацией.*

Elizarov A.M., Lipachev E.K., Khaydarov S.M.
Kazan (Volga Region) Federal University
Kazan, Russia
mailto: amelizarov@gmail.com, elipachev@gmail.com, 15jkeee@gmail.com

AUTOMATED SYSTEM OF STRUCTURAL AND SEMANTIC PROCESSING OF PHYSICAL AND MATHEMATICAL CONTENT

Abstract: Automatic processing service system of large collections of physical and mathematical documents presented. Services provided: automatic validation of documents and their conversion in accordance with the rules of formation; semantic analysis of documents, extraction of metadata, as well as the preparation of various types of scientific materials and publications with a choice of the further adjustment of their structure. The above services can automatically perform the processing of large collections of electronic documents a set of operations and activities that can not be realized with the traditional "manual" work with electronic content. One example of the successful use of the system is shown.

Keywords: *information extraction, data integration, information management systems.*

Введение. Современные электронные научные коллекции, такие, например, как архивы научных журналов и отчетов, сборники научных трудов, диссертации и др., являются составной частью электронных научных библиотек и представляют собой наборы документов, имеющих различную структуру и разные форматы представления текстовых и графических материалов, библиографических списков, математической нотации. Эти различия затрудняют организацию информационных сервисов, опирающихся на машиноориентированную обработку информации (см., например, [1, 2]). Кроме того, в настоящее время значительно увеличивается объем данных, включаемых в коллекции, что в свою очередь создает дополнительные трудности при обработке информации. Поэтому сейчас активно развиваются новые подходы, инструменты и методы обработки информации, а огромные объёмы обрабатываемых данных стали обозначать термином «большие данные» (Big Data). Одновременно все более востребованными у ученых становятся новые способы обнаружения объектов научного знания непосредственно через Веб, а также инструменты и сервисы, обеспечивающие создание и совместное использование новых видов структур знаний. В контексте концепции связанных данных (Linked Data) и Семантического Веба такие инструменты и сервисы можно использовать для создания графов сотрудничества, которые полезны, например, для вычисления «расстояния сотрудничества» (collaboration distance) между авторами и выделение «близких» документов, что открывает новые возможности тонкой настройки поиска и просмотра (см., например, [3–5]). Многими авторами (например, [6–9]) подчеркивается важность разработки новых онтологий предметных областей, в частности, в области математики, поскольку традиционной библиографической каталогизации сегодня уже недостаточно – требуется более глубокая детализация, содержащая описания, созданные с учетом разных точек зрения.

Проектом, нацеленным на реализацию и развитие названных новых подходов к обработке и использованию информации в области математики, стал проект организации всемирной Цифровой Математической Библиотеки (Digital Mathematics Library – DML) [6, 10]. Эта библиотека видится идеологам и разработчикам проекта как организационная структура, с помощью которой будут производиться агрегирование и осуществляться доступ к любой информации, являющейся значимой для математического сообщества. При этом функциональные возможности DML не должны ограничиваться простым предоставлением доступа к математическим публикациям – должны быть обеспечены возможности для аннотирования, поиска, просмотра, навигации, связывания, организации вычислений, визуализации любого контента, как защищенного авторским правом, так и открытого.

В базовых документах проекта подчеркнуто, что необходимо также интеллектуальное извлечение информации для последующей передачи пользователю [6, 8]. В качестве примера назван сервис, позволяющий пользователю выделить формулу, а затем обратиться к DML для получения разъяснений и необходимых ссылок.

Математикам необходима навигация по всему корпусу математических документов с возможностью их просмотра и получения дополнительной информации по интересующим объектам. Далее, помимо простого изучения цитирования работ, связанных с исследованием данного объекта, было бы полезным использовать информацию о том, что другие пользователи DML нашли интересного в связи с этим объектом, например, иметь возможность получать ответ на вопрос, какие статьи просмотрели читатели, которые также заинтересовались данными статьями. Так можно было бы найти документы, которые конкретно не ссылаются друг на друга, но относятся к одной и той же теме. В целом разработчики проекта DML полагают, что следующий шаг в продвижении математики состоит в выходе за пределы традиционных математических публикаций и построении сети информации, основанной на знаниях, содержащихся в этих публикациях.

Нами и нашими коллегами выполнен ряд исследований, лежащих в русле идеологии проекта DML (см. работы [3–5, 8, 9, 11–18]). Настоящая работа развивает эти результаты. В ней представлена система сервисов автоматической обработки больших коллекций физико-математических документов. Сервисы обеспечивают: автоматическую валидацию документов и их преобразование в соответствии с правилами формирования коллекций (в частности, правилами представления статей в научные журналы); семантический анализ документов, извлечение метаданных, а также подготовку различных типов изданий научных материалов с выбором и дальнейшей корректировкой их структуры. Названные сервисы позволяют автоматически выполнять при обработке больших коллекций электронных документов такой набор операций и действий, который не реализуем при традиционной «ручной» работе с электронным контентом, и включать полученные электронные коллекции в DML.

Система сервисов автоматической обработки коллекций научных документов. Машиноориентированная обработка электронных коллекций предполагает наличие семантической разметки их документов. Выполнить такую разметку можно в автоматическом режиме на основе информации о структурном строении каждого документа и особенностях его форматирования (см., например, [11–16, 19]). Коллекция разбивается на классы сходных по структуре документов, для каждого класса производится преобразование документов к семантическому представлению. С помощью набора паттернов регулярных выражений, специфичных для каждого класса документов, производится выделение информационных блоков (названия ста-

тьи, списка авторов, блока литературы и т. д.). В свою очередь, это дает возможность не только использовать семантические инструменты работы с электронным контентом, но и формировать в автоматическом режиме новые виды документов. Как пример, приведем систему сервисов, созданную нами для управления коллекцией материалов XI Всероссийского съезда по фундаментальным проблемам теоретической и прикладной механики (Казань, 20 – 24 августа 2015 г.). Эта система сервисов включает модули, выполняющие следующие функции:

- извлечение метаданных из документов коллекции на основе анализа их структуры и форматов представления информации [12, 16, 19, 20];
- автоматический выбор документов согласно установленному порядку, например, лексикографическому, по спискам авторов [16, 17];
- извлечение блоков аннотаций из документов коллекции, подготовка алфавитного указателя и формирование сборника аннотаций;
- автоматическое формирование библиографического описания статьи коллекции с записью этой информации в блок колонтитулов документа;
- конвертация документов в pdf-формат в соответствии с установленными параметрами;
- формирование сборника трудов Съезда с автоматической выборкой статей, расстановкой страниц, подготовкой алфавитного указателя и содержания;
- подготовка метаданных для экспорта в РИНЦ [21].

Заключение. Работа выполнена при финансовой поддержке РФФИ (проекты №№ 15-07-08522, 15-47-02472).

Литература

1. *Olver P.J.* Journals in flux // Notices Amer. Math. Soc. – 2011. – V. 58 (8). – P. 1124–1126.

2. *Афонин С.А., Бахтин А.В., Бухонов В.Ю., Васенин В.А., Ганкин Г.М., Гаспарянц А.Э., Голомазов Д.Д., Иткес А.А., Козицын А.С., Тумайкин И.Н., Шапченко К.А.* Интеллектуальная система тематического исследования научно-технической информации (ИСТИНА). Под ред. академика В.А. Садовничего. М.: Изд-во Московского ун-та, 2014. – 262 с.

3. *Елизаров А.М., Жильцов Н.Г., Иванов В.В., Кириллович А.В., Липачёв Е.К., Невзорова О.А.* Семантический рекомендательный сервис в профессиональной деятельности математика // Учёные записки Института социально-гуманитарных знаний. – 2015. – № 1. – С. 190–197.

4. *Елизаров А.М., Жильцов Н.Г., Кириллович А.В., Липачёв Е.К.* Семантическое аннотирование в системе управления физико-математическим контентом // Научный сервис в сети Интернет: труды XVII Всероссийской научной конференции (21–26 сентября 2015 г., г. Новороссийск). – М.: ИПМ им. М.В. Келдыша, 2015. – С. 98–103.

5. *Елизаров А.М., Жильцов Н.Г., Кириллович А.В., Липачёв Е.К.* Терминологическое аннотирование и рекомендательный сервис в системе управления физико-математическим контентом // Труды XVII Международной конференции DAMDID/RCDL'2015 «Аналитика и управление данными в областях с интенсивным использованием данных». Обнинск: ИАТЭ НИЯУ МИФИ, 2015. – С. 347–350.

6. Developing a 21st Century Global Library for Mathematics Research. Washington, The National Academies Press, 2014. – 131 p. URL: http://arxiv.org/abs/1404.1905.

7. *Staab S., Studer R. (Eds.)* Handbook on Ontologies. Berlin, Heidelberg: Springer Verlag, 2003, 2009. – 811 p.

8. *Elizarov A., Kirillovich A., Lipachev E., Nevzorova O., Solovyev V., Zhiltsov N.* Mathematical knowledge representation: semantic models and formalisms // Lobachevskii Journal of Mathematics. – 2014. – V. 35 (4). – P. 347–353.

9. *Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E.* OntoMathPro ontology: a linked data hub for mathematics // Communications in Computer and Information Science. – 2014. – V. 468. – P. 105–119.

10. *Olver P.J.* The world digital mathematics library: report of a panel discussion // Proceedings of the International Congress of Mathematicians, August 13–21, 2014, Seoul, Korea. Kyung Moon SA, 2014. – V. 1. – P. 773–785. URL: http://www.icm2014.org/download/download.asp?fn=Proceedings_Volume_I.pdf.

11. *Елизаров А.М., Липачёв Е.К., Хохлов Ю.Е.* Семантические методы структурирования математического контента, обеспечивающие расширенную поисковую функциональность// Информационное общество. – 2013. – № 1–2. – С. 83–92.

12. *Биряльцев Е.В., Елизаров А.М., Жильцов Н.Г., Липачёв Е.К., Невзорова О.А., Соловьев В.Д.* Методы анализа семантических данных математических электронных коллекций // Научно-техническая информация. Сер. 2. Информационные процессы и системы. – 2014. – № 4. – С. 12–17.

13. *Елизаров А.М., Липачёв Е.К., Невзорова О.А., Соловьев В.Д.* Методы и средства семантического структурирования электронных математических документов // Доклады Академии наук. – 2014. – Т. 457 (6). – С. 642–645.

14. *Elizarov A.M., Lipachev E.K., Zuev D.S.* mathematical content semantic markup methods and open scientific e-journals management systems // Communications in Computer and Information Science. – 2014. – V. 468. – P. 242–251.

15. *Елизаров А.М., Зуев Д.С., Липачёв Е.К.* Информационные системы управления электронными научными журналами // Научно-техническая информация. Сер. 1. Организация и методика информационной работы. – 2014. – № 3. – С. 31–38.

16. *Хайдаров Ш.М.* Семантический анализ документов в системе управления цифровыми научными коллекциями // Электронные библиотеки. – 2015. – Т. 18 (1–2). – С. 61–85.

17. *Хайдаров Ш.М.* Методы управления математическим контентом в информационных издательских системах // Тр. Матем. центра им. Н.И. Лобачевского. Материалы 14-й Всерос. Молодежной школы-конференции «Лобачевские чтения–2015 (Казань, 22–27 октября 2015 года). Казань, 2015. С. 162–165.

18. *Елизаров А.М., Жижченко А.Б., Жильцов Н.Г., Кириллович А.В., Липачёв Е.К.* Онтологии математического знания и рекомендательная система для коллекций физико-математических документов // Доклады Академии наук. – 2016. – Т. 467 (4). – С. 392–395.

19. *Tkaczyk D., Tarnawski B., Bolikowski L.* Structured affiliations extraction from scientific literature // D-Lib Magazine. – 2015. – V. 21 (11/12). URL: http://www.dlib.org/dlib/november15/tkaczyk/11tkaczyk.html.

20. Standard ECMA-376: Office Open XML File Formats. URL: http://www.ecmainternational.org/publications/standards/Ecma-376.htm.

21. *Герасимов А.Н., Елизаров А.М., Липачёв Е.К.* Формирование метаданных для международных баз цитирования в системе управления электронными научными журналами // Электронные библиотеки. – 2015. – Т. 18 (1–2). – С. 6–31.