Автоматизированная система сервисов обработки больших коллекций научных документов

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

Аннотация

Представлена система сервисов автоматической обработки коллекций научных документов. Эти сервисы обеспечивают проверку соответствия документов принятым правилам формирования коллекций и их преобразование в установленные форматы; структурный анализ документов и извлечение метаданных, а также их интеграцию в научное информационное пространство. Система позволяет автоматически выполнять набор операций, который не реализуем за практически приемлемое время при традиционной «ручной» обработке электронного контента, и предназначена для больших коллекций научных документов.

1 Введение

Сегодня одной из актуальных проблем, стоящих перед человечеством, стала проблема накопления и последующей обработки огромных массивов данных. Под данными традиционно подразумевают различные необработанные информационные материалы, в том числе, данные различных наблюдений и научных экспериментов, персональные данные, а также различную статистическую информацию. По сведениям, приведенным в [1], уже в 2011 году каждый день создавалось около 15 PB новых данных, а за три года до этого момента времени человечество произвело информации больше, чем за всю историю своего существования до 2008 года, причем прирост данных происходил экспоненциально: это были и научные данные, и сведения о проведенных операциях-транзакциях, отчеты в социальных сетях и многое другое. Сегодня мировой объем данных увеличивается более чем в два раза каждые два года, а большие объемы данных (которые с 2008 года стали обозначать термином «большие данные» (Big Data)) открывают новые возможности и существенно влияют на развитие информационно-коммуникационных технологий (ИКТ).

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

Традиционно термином Big Data обозначают наборы данных таких объёма и сложности, что стандартные инструменты работы с данными не способны осуществлять их обработку за время, приемлемое для практики [2]. Более широко этот термин можно трактовать как набор эффективных подходов, методов и инструментов обработки различных структурированных и неструктурированных данных большого объёма с целью получения приемлемых результатов в условиях непрерывного прироста данных [3]. Другими словами, термин «большие данные» характеризует совокупности данных, которые слишком велики по объему, характеризуются экспоненциальным ростом, не форматированы или не структурированы для анализа традиционными методами.

Не менее актуальна проблема учета значительного роста объёмов данных, получаемых, хранимых и обрабатываемых в ходе научной деятельности. В настоящее время благодаря внедрению ИКТ в научно-исследовательскую деятельность стало возможным при проведении новых исследований использовать весь корпус накопленных научных знаний. Это предполагает создание комплекса технологий, обеспечивающих оптимальное управление имеющимися знаниями, организацию эффективного доступа к ним, а также совместное и многократное использование новых видов структур знаний. В результате формируются разнообразные электронные научные коллекции и библиотеки, такие, например, как архивы научных журналов и отчетов, сборники научных трудов, диссертации и др. Они 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, [4, 5]). Кроме того, в настоящее время значительно увеличивается объем данных, включаемых в коллекции, что в свою очередь создает дополнительные трудности при обработке научных Big Data. При управлении электронными научными коллекциями больших данных в полной мере остаются актуальными, а также появляются новые задачи, в их числе: се-

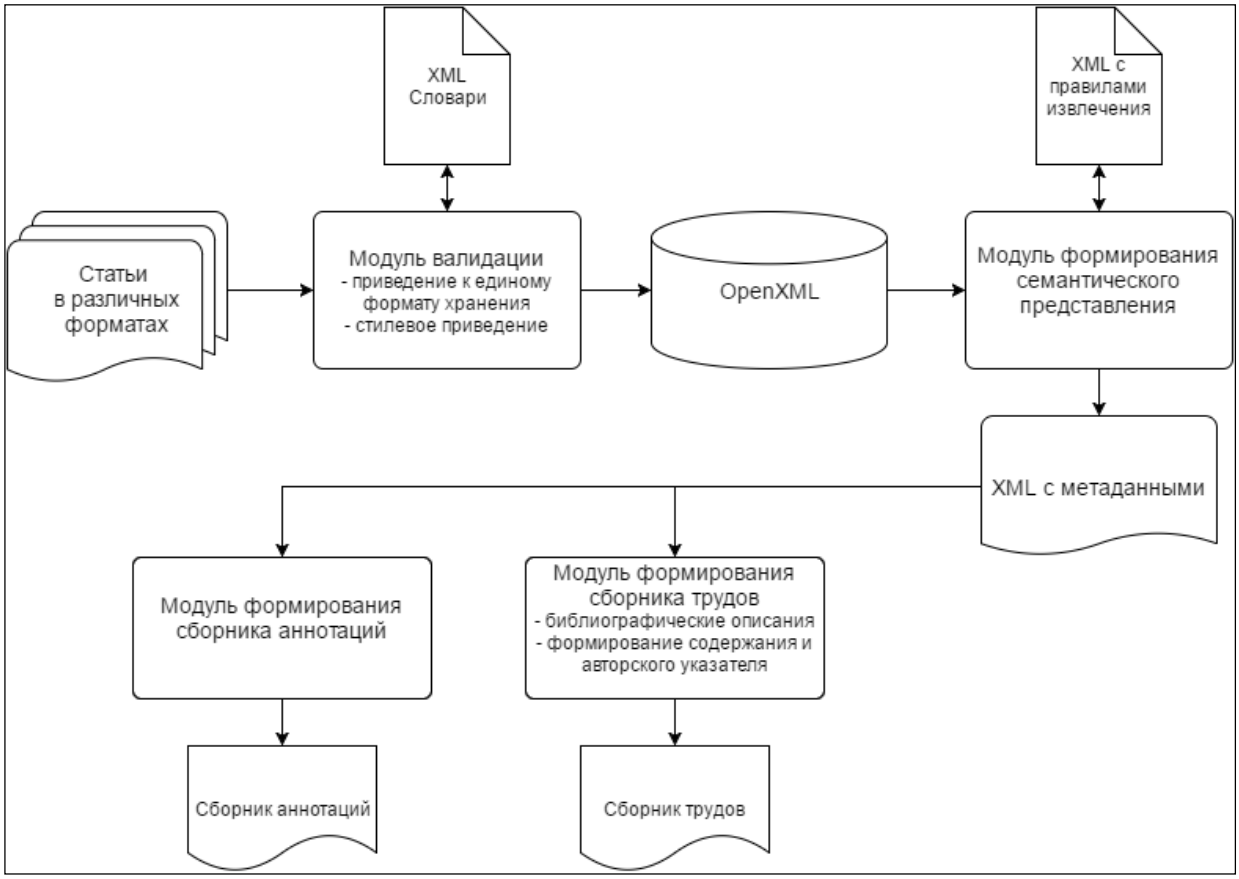

Рис. 1. Архитектура системы

мантическая разметка, организация поиска, выделение метаданных, формирование тематических кластеров документов, сбор наукометрической информации, подготовка сборников материалов и др. (см., например, [6, 7]). Насущными становятся проблемы анализа и управления данными в различных областях с интенсивным использованием данных. Ниже представлена система сервисов автоматической обработки коллекций научных документов. С ее использованием проведена обработка материалов XI Всероссийского съезда по фундаментальным проблемам теоретической и прикладной механике (далее – Съезд), проведенного в Казани 20 – 24 августа 2015 г.: сформированы программа съезда, сборники аннотаций и трудов съезда (объемом более 1500 статей), а также соответствующая электронная коллекция.

2 Архитектура системы сервисов

На рис. 1 представлена архитектура созданной автоматизированной системы сервисов обработки больших коллекций научных документов. Она состоит из модулей, выполняющих следующие функции:

• извлечение метаданных из документов коллекции на основе анализа их структуры и форматов представления информации;

• автоматический выбор документов согласно установленному порядку, например, лексикографическому, по спискам авторов;

• 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 оригинал-макетов планируемых изданий с автоматической выборкой статей, расстановкой страниц, подготовкой алфавитного указателя и содержания;

• подготовка метаданных для экспорта в базы данных Российского индекса научного цитирования (РИНЦ).

3 Организация электронного хранилища

Машиноориентированная обработка электронных коллекций предполагает наличие семантической разметки их документов. Такая разметка частично присутствует в документах, использующих TEX-нотацию, при условии, что используются соответствующие макрокоманды (например, \title, \author, \abstract, \keywords) и стилевое окружение,

характерное для каждой коллекции. В электронных научных коллекциях, представленных в офисных форматах (.doc, .docx и др.), а также .pdf, семантическая разметка отсутствует. Тем не менее, выполнить такую разметку можно в автоматическом режиме на основе информации о структурном строении каждого документа и особенностях его форматирования.

Прежде всего, 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 статьи, 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.

В хранилище организована навигация по названию, авторам и т. д. Реализация этих сервисов основана на структурном анализе документов в коллекции (см. раздел 5).

4 Сервис валидации и стилевого приведения

Под валидацией документов коллекции понимается процесс проверки наличия и расположения ключевых блоков (название статьи, список авторов, аффилиация, ключевые слова и т. д.), указанных в регламентируемых документах.

Сервис стилевого приведения реализует следующие шаги:

- единообразное представление названий статей; списка авторов (например, вместо Хайдаров Ш.М. записывается Ш.М. Хайдаров);
- единообразное представление аффилиации авторов, например, записи «КФУ», «К(П)ФУ», «Казанский университет», «Казанский (Приволжский) федеральный университет» и «Казанский федеральный университет» приводятся к единому виду «Казанский (Приволжский) федеральный университет»; для этого создается словарь синонимов;
- единообразное шрифтовое оформление разделов текста статей; происходит учет регистра при записи ключевых блоков, например, название статьи записывается прописными буквами;
- осуществляется выбор форматов рисунков, схем, диаграмм;
- производится набор математических формул и системы ссылок на них;
- списки литературы приводятся к выбранному формату библиографического описания.
- оформляются ссылки на поддержку исследований грантами, благодарности.

5 Формирование семантического представления коллекции на основе структурного анализа

Для извлечения метаданных статьи по характерным признакам (см. таблицу 1) определяются правила выделения блоков статьи. К таким признакам относятся стилевое оформление статей (шрифт, размер шрифта, выделение и т. д.). Кроме того, такие дополнительные признаки, как шаблонность текста (например, слово «Аннотация» перед блоком аннотаций или шаблонный вид электронной почты) и положение блока в тексте (например, документ начинается с названия статьи), позволяют повысить качество извлечения. В качестве таких признаков могут использоваться положение блока в документе, а также шрифт используемый в данном блоке (см., например, [8–11]). При структурном анализе коллекции научных документов Съезда использовался набор признаков, указанный в таблице 1.

Блок статьи	Признаки блока
Название статьи	Шрифт: Times New Roman, 12 пт, полужирный, выравнивание по центру. Положение: в начале документа
Список авторов	Шрифт: Times New Roman, 12 пт, выравнивание по центру Положение: после названия Шаблон имеют вид: И.О. Фамилия или И. Фамилия, перечисляются через запятую
Аффилиация	Шрифт: Times New Roman, 12 пт, курсив, выравнивание по центру. Положение: после списка авторов.
Электронная почта	Шрифт: Times New Roman, 9 пт, выравнивание по центру Положение: после аффилиации Шаблон содержат символ @ и имеют заданный вид
Аннотация	Шрифт: Times New Roman, 9 пт, выравнивание по ширине Положение: после адреса электронной почты Шаблон начинается со слова «Аннотация».

Таблица 1. Характерные признаки для извлечения метаданных

Модуль реализован в виде PHP-скрипта, и его работа состоит из следующих шагов. Из файла статьи хранящемся в формате docx, извлекается файл document.xml (см., например, [12]). Далее с использованием описания класса DOMDocument производится разбор этого файла. Для выделения блоков применяется метод getElementsByTagNameNS с параметром «w:p» (тег разметки абзаца в OpenXML). В результате получается список всех абзацев документа как объекта DOMNodeList. Полученный список последовательно проверяется на соответствие заданным правилам. В итоге для каждого документа (см. пример рис. 2) формируется его семантическое представление (рис. 3).

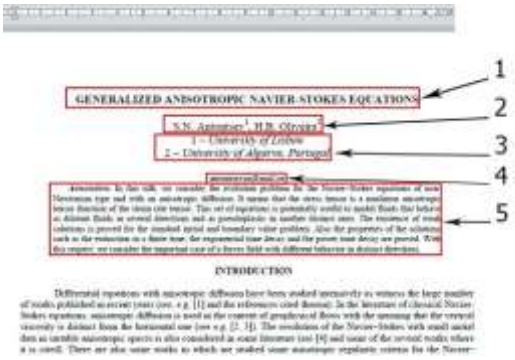

Рис. 2. Пример статьи, где 1 – название, 2 – блок списка авторов, 3 – блок аффилиации, 4 – блок электронной почты и 5 – блок аннотации. Стилевое оформлении соответствует таблице 1

```
<?xml version="1.0" encoding="UTF-8"?>
<articles>
 <article>
 <eachauthors>
  <author>S.N. Antontsev</author>
  <author>H.B. Oliveira</author>
  <workplace>University of Lisbon; University of Algarve, Portugal</workplace>
  <mails>antontsevsn@mail.ru</mails>
 </eachauthors>
 <artTitles>
  <artTitle>GENERALIZED ANISOTROPIC NAVIER-STOKES EQUATIONS</artTitle>
 </artTitles>
 <abstracts>
 <abstract>&lt;p style=&quot;text-indent: 20px;&quot;&gt;In this talk, we consider the evolution problem for the Navier–Stokes equations of non-Newtonian type and with an anisotropic diffusion.
                    ...
&lt;/p&gt;</abstract>
 </abstracts>
 <files>
  <furl>F:\Desktop\doc/00001.pdf</furl>
 </files>
</article>
…
</articles>
```

Рис. 3. Фрагмент сгенерированного XML-файла

Для выделения семантических элементов разработан набор регулярных выражений, например, для выделения списка авторов используется выражение

/([А-ЯА-Z]\.(?:[А-ЯА-Z]\.)?\s[А-ЯА-Z][а-zа-я]+)(,\s)?(?1)?(,\s)?(?1)?/

Кроме того, проверяются наличие ключевых конструкций и их соответствие заданному формату. Результатом работы описываемого модуля является XML-документ, содержащий метаданные статей размечаемой коллекции.

6 Модуль формирования оригинал-макета научного издания

Этот модуль позволяет в автоматическом режиме подготовить из файлов электронной коллекции оригинал-макет научного издания (сборник материалов, труды и т. д.). Порядок размещения статей определяется семантическим представлением коллекции, хранящемся в XML-файле (см. раздел 5). Алгоритм реализован в виде макроса VBA и включает следующие шаги: сначала для задания диапазона страниц статей определяются счетчики начальной и конечной страниц и задаются их начальные значения (см. рис. 4–6). Далее последовательно открываются документы коллекции в соответствии с порядком, заданным в XML-файле в соответствии с правилами извлечения. Вычисляются начальные и конечные страницы, после чего формируется библиографическое описание статьи, которое записывается в колонтитул данного документа. Полученный документ конвертируется в PDF-формат. Также библиографическое описание сохраняется в XML-файле. На рис. 6 приведен фрагмент кода, выполняющий описанные операции.


ТЕОРИЯ ВАРИАЦИОННЫХ ОБРАТНЫХ КРАЕВЫХ ЗАДАЧ АЭРОГИДРОДИНАМИКИ: СОВРЕМЕННОЕ СОСТОЯНИЕ, ПРИЛОЖЕНИЯ, ПЕРСПЕКТИВЫ РАЗВИТИЯ

А.М. Елизаров

Казанский (Приволжский) федеральный университет

ОСНОВНЫЕ РЕЗУЛЬТАТЫ ТЕОРИИ ВАРИАЦИОННЫХ ОКЗА

Рис. 4. Фрагмент документа до обработки модулем


ТЕОРИЯ ВАРИАЦИОННЫХ ОБРАТНЫХ КРАЕВЫХ ЗАДАЧ АЭРОГИДРОДИНАМИКИ: СОВРЕМЕННОЕ СОСТОЯНИЕ, ПРИЛОЖЕНИЯ, ПЕРСПЕКТИВЫ РАЗВИТИЯ

А.М. Елизаров

Казанский (Приволжский) федеральный университет

ОСНОВНЫЕ РЕЗУЛЬТАТЫ ТЕОРИИ ВАРИАЦИОННЫХ ОКЗА

Рис. 5. Фрагмент документа после обработки модулем (создан колонтитул с выходными данными)

```
Sub Макрос1()
'
' Макрос1 Макрос
'
'
 Application.ScreenUpdating = False
 StartPage = 4367
 sPath = "F:\Desktop\doc1\"
 sFile = Dir(sPath & "*.docx")
 While sFile <> ""
  With Documents.Open(sPath & sFile)
  ActiveDocument.PageSetup.HeaderDistance = Centimeter-sToPoints(1)
  deleteAllHeaders_Footers
  ActiveDocu-ment.PageSetup.DifferentFirstPageHeaderFooter = True
  ActiveWindow.ActivePane.View.SeekView = wdSeekFirstPageHeader
  ActiveDocument.Repaginate
  EndPage = StartPage + ActiveDocu-ment.BuiltInDocumentProperties(wdPropertyPages) - 1
  Selection.Font.Name = "Times New Roman"
  Selection.Font.Size = 9
  Selection.Font.Bold = wdToggle
  Selection.Font.Italic = wdToggle
  Selection.TypeText Text:="XI Всероссийский съезд по фундаментальным проблемам теоретической и прикладной механики,"
  Selection.TypeParagraph
  Selection.TypeText Text:="Казань, 20 – 24 авгу-ста 2015 года. С. "
  Selection.TypeText StartPage
  Selection.TypeText Text:="-"
  Selection.TypeText EndPage
  Selection.TypeText Text:="."
  Selection.TypeParagraph
  Selection.InlineShapes.AddHorizontalLineStandard
  Selection.MoveLeft Unit:=wdCharacter, Count:=2, Ex-tend:=wdExtend
  Selection.InlineShapes(1).Fill.Visible = msoTrue
  Selection.InlineShapes(1).Fill.Solid
  Selection.InlineShapes(1).Fill.ForeColor.RGB = RGB(0, 0, 0)
  Selection.InlineShapes(1).Fill.Transparency = 0#
  Selec-tion.InlineShapes(1).HorizontalLineFormat.WidthType = _
   wdHorizontalLinePercentWidth
  Selec-tion.InlineShapes(1).HorizontalLineFormat.PercentWidth = 100
  Selection.InlineShapes(1).Height = 1
  Selec-tion.InlineShapes(1).HorizontalLineFormat.NoShade = True
  Selection.InlineShapes(1).HorizontalLineFormat.Alignment = _
   wdHorizontalLineAlignCenter
  ActiveWindow.ActivePane.View.SeekView = wdSeekMainDoc-ument
  ActiveWindow.ActivePane.View.SeekView = wdSeekFirstPage-Footer
  Selection.Fields.Add Range:=Selection.Range, Type:=wdFieldEmpty, PreserveFormatting:=False
  Selection.TypeText Text:="PAGE"
  Selection.Fields.Update
  Selection.Fields.ToggleShowCodes
  Selection.Font.Name = "Times New Roman"
  Selection.Font.Size = 12
  Selection.ParagraphFormat.Alignment = wdAlign-ParagraphRight
  ActiveWindow.ActivePane.View.SeekView = wdSeekMainDoc-ument
  With ActiveDocu-ment.Sections(1).Footers(wdHeaderFooterPrimary).PageNumbers
   .RestartNumberingAtSection = True
   .StartingNumber = StartPage
   .Add wdAlignPageNumberRight, False
  End With
  StartPage = EndPage + 1
  ActiveDocument.ExportAsFixedFormat Output-FileName:=sPath & Replace(sFile, "docx", "pdf"), Export-Format _
  :=wdExportFormatPDF, OpenAfterExport:=False, OptimizeFor:= _
  wdExportOptimizeForPrint, Range:=wdExportAllDocument, From:=1, To:=1, _
  Item:=wdExportDocumentContent, Includ-eDocProps:=True, KeepIRM:=True, _
  CreateBookmarks:=wdExportCreateNoBookmarks, DocStruc-tureTags:=True, _
  BitmapMissingFonts:=True, UseI-SO19005_1:=False
  On Error Resume Next
  ActiveDocument.Close (True)
  End With
  sFile = Dir
 Wend
End Sub
```

Рис. 6. Фрагмент кода формирования библиографического описания


XI Всероссийский съезд по фундаментальным проблемам теоретической и прикладной механики, Казань, 20 – 24 августа 2015 года. С. 4429-4479.

СОДЕРЖАНИЕ

Рис. 7. Пример автоматически сгенерированного содержания сборника трудов Съезда

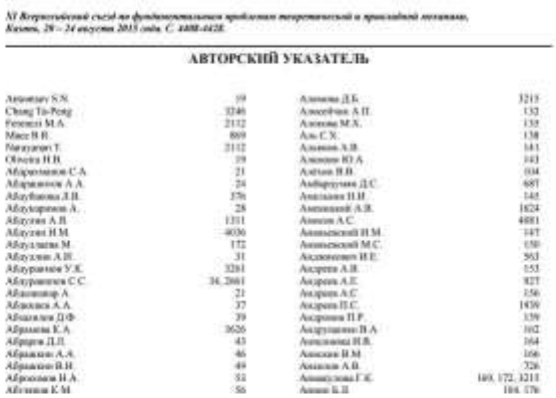
АВТОРСКИЙ УКАЗАТЕЛЬ

Рис. 8. Пример автоматически сгенерированного авторского указателя сборника трудов Съезда

После обработки всех документов коллекции формируются содержание издания и авторский указатель. При этом используются данные, сохраненные в XML-файле на этапе формирования колонтитулов. На рисунках 7 и 8 приведены автоматически сформированные содержание издания и авторский указатель.

7 Сервис извлечения библиографических метаданных и загрузки в РИНЦ

Алгоритм извлечения библиографических метаданных и загрузки их в РИНЦ состоит из следующих шагов (проиллюстрированных на примере материалов Съезда):

- из оригинал-макета сборника трудов извлечены библиографические описания каждой публикации;
- соответствующий скрипт находит в документе блок библиографических описаний и с помощью регулярных выражений разделяет их по видам изданий (например, отличительным признаком библиографического описания статьи является наличие знака //);
- проводится разбор основных метаданных – выделяются список авторов, названия статей, изданий и т. д.;
- с помощью разработанного веб-приложения генерируется XML-файл в соответствии с правилами РИНЦ, содержащий набор метаданных публикации.

Заключение

Предложен метод автоматической обработки больших коллекций физико-математических документов, включающий их валидацию и семантический анализ, извлечение метаданных, подготовку различных видов оригинал-макетов научных изданий. Метод позволяет выполнять автоматическую обработку больших коллекций электронных документов с набором операций, который не реализуем при традиционной «ручной» работе с электронным контентом.

Приведен пример успешной его реализации при организации XI Всероссийского съезда по фундаментальным проблемам теоретической и прикладной механики (Казань, 20 – 24 августа 2015 г.).

Благодарности

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

Литература

[1] IBM’s Top Storage Predictions for 2011, January 2011, StorageNewsletter.com. **Ошибка! Недопустимый объект гиперссылки.**

[2] MIKE2.0. The open source standard for information management. Big Data definition. http://mike2.openmethodology.org/wiki/Big_Data_Definition

[3] A Manyika J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H. Byers. Big data: The next frontier for innovation, competition, and productivity: McKinsey Global Institute Report, 2011. http://www.mckinsey.com/insights/mgi/ research/technology_and_innovation/big_data_the_ next_frontier_for_innovation

[4] P. J. Olver. Journals in flux. Notices Amer. Math. Soc., V. 58 (8), 2011, p. 1124-1126.

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

[6] А. М. Елизаров, Н. Г. Жильцов, А. В. Кириллович, Е. К. Липачёв. Семантическое аннотирование в системе управления физико-математическим контентом. Науч. сервис в сети Интернет: труды XVII Всерос. науч. конф. (21–26 сентября 2015 г., г. Новороссийск), М.: ИПМ им. М.В. Келдыша, с. 98-103, 2015.

[7] А. М. Елизаров, Н. Г. Жильцов, А. В. Кириллович, Е. К. Липачёв. Терминологическое аннотирование и рекомендательный сервис в системе управления физико-математическим контентом. Труды XVII Межд. конф. DAMDID / RCDL’2015 «Аналитика и управление данными в областях с интенсивным использованием данных». Обнинск: ИАТЭ НИЯУ МИФИ, с. 347-350, 2015.

[8] Xiaonan Lu, Brewster Kahle, James Z. Wang and C. Lee Giles. A metadata generation system for scanned scientific volumes. Joint Conference on Digital Libraries, June 16–20, 2008, Pittsburgh, Pennsylvania, p. 167-176, 2008. **Ошибка! Недопустимый объект гиперссылки.**

[9] J. Chen, H. Chen. A structured information extraction algorithm for scientific papers based on feature rules learning. Journal of Software, Vol. 8(1), p. 55-62, 2013. http://www.jsoftware.us/vol8/ jsw0801-08.pdf

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

[11] А.М. Елизаров, Е.К. Липачёв, Ш.М. Хайдаров. Автоматизированная система структурной и семантической обработки физико-математического контента. Ученые записки Института социально-гуманитарных знаний, № 1 (14), с. 210-215, 2016.

[12] Standard ECMA-376: Office Open XML File Formats. http://www.ecma-international.org/ publications/standards/Ecma-376.htm

Automatic processing service system of large collections of scientific documents

Alexander M. Elizarov, Evgeny K. Lipachev,
Shamil M. Khaydarov

This paper presents a system of automatic processing of scientific documents collection services. These services provide verification of compliance, document accepted rules of formation of collections and their conversion to established formats; structural analysis of documents and extraction of metadata, as well as their integration into the scientific information space. The system allows you to automatically perform a set of operations that cannot be realized for practical time with the traditional simple manual handling processing of electronic content. It is designed for large collections of scientific documents.