Методы автоматизированного извлечения метаданных научных публикаций для библиографических и реферативных баз цитирования

А.Н. Герасимов, А.М. Елизаров, Е.К. Липачев, Ш.М. Хайдаров
Казанский (Приволжский) федеральный университет
sav241@mail.ru, amelizarov@gmail.com,
elipachev@gmail.com, 15jkeee@gmail.com

Аннотация

Предложен алгоритм автоматического извлечения библиографических данных из однородного массива публикаций (в частности, выпусков научного журнала) и формирования блоков метаданных для экспорта в международные информационно-аналитические системы.

Ключевые слова: издательские системы; электронный научный журнал; интеграция электронных ресурсов; базы данных научного цитирования; экстракция метаданных

1. Введение

В соответствии с международными стандартами подготовка к публикации (в том числе, в электронной форме) любого научного журнала (текущего номера или выпуска с соответствующим набором статей) предполагает, в частности, выполнение ряда требований, выдвигаемых индексами научного цитирования, формируемыми в автоматизированном режиме (см. [1, 2]). К последним, например, относятся базы данных Scopus, Web of Science и Российского индекса научного цитирования (РИНЦ), получившие за последнее десятилетие широкое распространение в научном мире и активно используемые для оценки научного уровня как самих журналов, так и публикуемых ими статей с помощью целого набора показателей (различные импакт-факторы журналов, индексы цитирования и др.). Формирование и обработка таких показателей сегодня невозможны без применения специализированных компьютерных систем. Эти системы используют особый вид информационных ресурсов, называемых метаданными (см., например, [3]). Как правило, набор метаданных научной публикации включает библиографическое описание статьи (авторы, название, источник (например, журнал), год издания, том, номер, начальная и конечная страницы), авторское резюме (аннотация, реферат) и ключевые слова, названия и места расположения организаций, от имени которых авторы представили свои материалы; фамилии ученых (членов редколлегий или рецензентов), представивших статью к публикации. В настоящее время актуальной стала

подготовка метаданных в автоматизированном режиме, особенно в связи с наблюдаемым ростом объемов создаваемой научной информации [4]. Отметим, что указанные компьютерные системы используют специальные форматы представления метаданных, что требует дополнительной работы с данными.

2. Методы формирования метаданных

В настоящее время существует множество научно-информационных систем, которые предназначены для подготовки и выпуска электронных журналов и позволяют управлять основными издательскими процессами: приёмом и рецензированием материалов, подготовкой номеров журнала и созданием метаданных [5, 6] (при этом часть метаданных формируется на основе информации, предоставляемой авторами при подаче статей). Особенно актуальна задача автоматизации обработки больших объемов информации [7, 8, 9].

Примером развитой научно-информационной системы является Open Journal Systems (OJS) (сравнение имеющихся систем проведено в [10]). OJS обладает модулями экспорта метаданных в форматах XML (по шаблону native.dtd), Erudit (в виде DTD), CrossRef XML, PubMed XML и др. (см. [11, 12]). Однако использование таких модулей становится нецелесообразным, когда имеющиеся архивы или коллекции были сформированы по технологиям, не согласованным с требованиями OJS. Такая ситуация возникла при подготовке и проведении 20 – 24 августа 2015 года в Казани XI Всероссийского съезда по фундаментальным проблемам теоретической и прикладной механики: в частности, потребовалось решить задачу автоматизированной подготовки метаданных публикаций (в соответствии с правилами баз научного цитирования) общим объемом более 1500 статей в формате .docx и загрузки в базы данных РИНЦ. Решение этой задачи основано на анализе структуры документов и учете стилевых правил оформления представленных материалов.

3. Автоматизация извлечения метаданных из коллекций научных документов

Основные этапы извлечения метаданных описаны ниже.

1. В автоматическом режиме был сформирован оригинал-макета сборника трудов съезда, из которого были извлечены библиографические описания (авторы, название, год издания, том, номер, начальная и конечная страницы) и ключевые слова каждой публикации (рис. 1); для этого был использован соответствующий скрипт (рис. 2).

XI Всероссийский съезд по фундаментальным проблемам теоретической и прикладной механики, Казань, 20 – 24 августа 2015 года. С. 1279-1281.

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

А.М. Елизаров
Казанский (Приволжский) федеральный университет

amelizarov@gmail.com

Аннотация. Вариационные обратные краевые задачи аэрогидродинамики (ОКЗА) реализуют один из подходов к оптимизации аэродинамических и гидродинамических форм, в частности, они связаны с поиском ответа на вопросы, какую максимальную подъемную силу можно получить на профиле крыла и какова форма профилей, обладающих оптимизированными аэродинамическими характеристиками. В рамках классических моделей механики жидкости и газа в математическом плане эти задачи сводятся к вариационным краевым задачам для аналитических функций.

Рис. 1. Первая страница публикации, содержащая основные метаданные

```
    //Номера страниц из файла колонтитулов
    $zip->extractTo('./tmp', array('word/header1.xml'));
    if(file_exists('tmp/word/header1.xml')){
        $xml->load('tmp/word/header1.xml');
        $header3=dnl2str($xml-
>getElementsByTagNameNS('http://schemas.openxmlformats.org/
        /wordprocessingml/2006/main','p'));
        unlink('tmp/word/header1.xml');
    }
    …
    $w_ps=$xml->getElementsByTagNameNS('http://schemas.openxmlformats.org
/wordprocessingml/2006/main','p');
    //Регулярное выражение, обрабатывающее фио
    $fio="/([А-ЯA-Z]\.(?:[А-ЯA-Z]\.)?\s[А-ЯA-Z][a-zа-я]+)(,\s)?(?1)?(,\s)?(?1)?/u";
    //Название статьи. Цикл чтоб обрабатывать отступы в названии статей.
    $articlename=$w_ps->item($k)->nodeValue;
    while(($w_ps->item($k+1)->nodeValue)!=""){
      if(preg_match($fio,$w_ps->item($k+1)->nodeValue))break;
      $articlename.=$w_ps->item($k+1)->nodeValue."%%";
      $k++;
    }
    $articlename=trim(preg_replace("/%%/u"," ",trim(mb_strtoupper($articlename,
'UTF-8'),"% ")." "));
    //Авторы. Место работы. E-mail.
    $mail_reg="/[-a-z0-9!#$%&'*+\/=?^_`{|}~]+(?:\.[-a-z0-
9!#$%&'*+\/=?^_`{|}~]+)*@(?:[a-z0-9]
     ([-a-z0-9]{0,61}[a-z0-
9])?\.)*(?:aero|arpa|asia|biz|cat|com|coop|edu|gov|info|int|jobs|mil|mobi|museum
      |name|net|org|pro|tel|travel|[a-z][a-z])/iu";
    while($k<$w_ps->length){
      if(preg_match($fio,$w_ps->item($k)->nodeValue)){
```

```
        if(substr(trim($author),-1)==','||!$author)
          $author.=$w_ps->item($k)->nodeValue;
        }
        if(preg_match($mail_reg,$w_ps->item($k)->nodeValue)){
          $mails.=$w_ps->item($k)->nodeValue;
        }
        }
        if(preg_match($fio,$w_ps->item($k)-
>nodeValue)&&!$workplace&&!$mails){
          $curwp=$k+1;
          while($w_ps->item($curwp)->getElementsByTagName('i')->length &&
                                  $w_ps->item($curwp)->nodeValue!=''){
            $workplace.=trim($w_ps->item($curwp)->nodeValue)."%";
            $curwp++;
          }
        }
        $k++;
      }
```

Рис. 2. Фрагмент скрипта извлечения библиографических описаний

2. Далее был выделен и обработан блок литературы (рис. 3). С использованием регулярных выражений этот блок был разделен на отдельные элементы — библиографические списки статей. Фрагмент соответствующего кода приведен на рис. 4.

Литература

1. Лаврентьев М.А. // Труды ЦАГИ. 1934. Вып. 155. 47 с.
2. Елизаров А.М., Ильинский Н.Б., Поташев А.В.// Изв. АН СССР. МЖГ. 1988. № 3. С. 5-13.
3. Елизаров А.М., Ильинский Н.Б., Поташев А.В. Обратные краевые задачи аэрогидродинамики. Итоги науки и техники. Сер. Механика жидкости и газа. М.: ВИНИТИ, 1989. Т. 23. С. 3-115.
4. Елизаров А.М., Федоров Е.В.// ПММ. 1990. Т. 54 (4). С. 571-580.
5. Елизаров А.М., Фокин Д.А. // Изв. АН СССР. МЖГ. 1990. № 3. С. 157-164.
6. Елизаров А.М., Ильинский Н.Б., Поташев А.В. Обратные краевые задачи аэрогидродинамики: теория и методы проектирования и оптимизации формы крыловых профилей. М.: Физматлит, 1994. 436 с.
7. Avkhadiev F.G., Elizarov A.M., Fokin D.A.// Europeen J. of Appl. Math. 1995. V. 6 (5). P. 385-398.

Рис. 3. Оформление библиографического списка статьи в сборнике трудов съезда

```
    //извлекаем document.xml из каждого docx файла в коллекции.
      foreach($files as $file){
        $zip = new ZipArchive;
        $res = $zip->open($file);
        if ($res === TRUE) {
          $zip->extractTo('./tmp', array('word/document.xml'));
          $zip->close();
          $file=iconv("cp1251","UTF-8", $file);
        } else {
```

```
            echo "$file failed to extract.<br>";
        }
        //echo $file.'<br/>';
        $file=str_replace("docx", "pdf",$file);

        //обрабатываем document.xml
        $xml = new DOMDocument;
        $xml->load('tmp/word/document.xml')
        $w_ps=$xml->getElementsByTagNameNS(
          'http://schemas.openxmlformats.org/wordprocessingml/2006/main','p');
       //цикл по абзацам документа
       while($k<$w_ps->length){
        ...
        //Поиск библиографии
        if(preg_match("[0-9](.*?)//(.*?)(\d{4})(.*?)(P.|C.)(.*?)(\.)" ,
              $w_ps->item($k)->nodeValue )){
        }
```

Рис. 4. Фрагмент скрипта выделения библиографического списка

3. После этапа формирования метаданных в автоматическом режиме были выделены тексты статей в соответствии с правилами загрузки в РИНЦ.

4. Разработанное веб-приложение допускает настройку формата экспорта метаданных, что позволяет задать необходимую структуру результирующего файла. С помощью веб-приложения (рис. 5) был сгенерирован XML-файл (рис. 6), записанный в соответствии с правилами РИНЦ и содержащий набор метаданных публикации.

```
    foreach ($articles as $article) {
    $xml .= <<<__ENDXML
      <article>
    <pages>{$article['pages']}</</pages>
    <artType>{$article['type']}</</artType>  __ENDXML;
       foreach ($article['authors'] as $author) {
        $xml .= <<<__ENDXML
          <author >
           <individInfo lang="{$author ['aut_lang']}">
           <surname>{$article['aut_sur_name']} </surname>
           <initials>{$article['aut_io']} </initials>
           <orgName>{$article['aut_org']} </orgName>
           <email>{$article['aut_email']} </email>
           </individInfo>
          </author>
    __ENDXML;
}
```

Рис. 5. Фрагмент PHP-кода, формирующего блок статьи в XML файле

```
<article>
<pages>1279-1281</pages>
<artType>PRC</artType>
<authors>
<author num="" id="">
<individInfo lang="RUS">
<surname>Елизаров</surname>
<initials>A.M.</initials>
<orgName>Казанский (Приволжский) федеральный университет</orgName>
<email>amelizarov@gmail.com</email>
</individInfo>
</author>
</authors>
<artTitles>
<artTitle lang="RUS">ТЕОРИЯ ВАРИАЦИОННЫХ ОБРАТНЫХ КРАЕВЫХ ЗАДАЧ
АЭРОГИДРОДИНАМИКИ: СОВРЕМЕННОЕ СОСТОЯНИЕ, ПЕРСПЕКТИВЫ РАЗВИТИЯ</artTitle>
<artTitle lang="ENG">THE THEORY OF VARIATIONAL INVERSE BOUNDARY VALUE PROBLEMS
AERODYNAMICS: CURRENT STATE AND PERSPECTIVES OF DEVELOPMENT</artTitle>
</artTitles>
<abstracts>
<abstract lang="RUS">Введение…..</abstract>
</abstracts>
<text lang="RUS">Полные тексты статей…..</text>
<references>
<reference>Лаврентьев М.А. // Труды ЦАГИ. 1934. Вып. 155. 47 с.</reference>
<reference>Елизаров А.М., Ильинский Н.Б., Поташев А.В.// Изв. АН СССР. МЖГ. 1988. No 3. С. 5-13.</reference>
<reference>Елизаров А.М., Ильинский Н.Б., Поташев А.В. Обратные краевые задачи аэрогидродинамики. Итоги науки и техники. Сер. Механика жидкости и газа. М.: ВИНИТИ, 1989. Т. 23. С. 3-115.</reference>
<reference>Елизаров А.М., Федоров Е.В.// ПММ. 1990. Т. 54 (4). С. 571-580.</reference>
<reference>Елизаров А.М., Фокин Д.А. // Изв. АН СССР. МЖГ. 1990. No 3. С. 157-164.</reference>
<reference>Елизаров А.М., Ильинский Н.Б., Поташев А.В. Обратные краевые задачи аэрогидродинамики: теория и методы проектирования и оптимизации формы крыловых профилей. М.: Физматлит, 1994. 436 с.</reference>
```

Рис. 6. Фрагмент XML-файла, содержащего метаданные статьи

4. Заключение

Эффективность предложенного алгоритма была подтверждена результатами обработки массива статей, представленных на съезд: из 1523 документов только 19 потребовали дополнительной ручной обработки.

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проекты 15-07-08522, 15-47-02472) и Российского гуманитарного научного фонда (проект 14-03-12004).

Литература

[1] Кириллова О.В. О системе включения журналов в БД Scopus: основные требования и порядок представления. URL: http://www.webcitation.org/68vOlqztg.

[2] Кириллова О.В. Критерии отбора и рекомендации по подготовке журнала в индекс цитирования Scopus. URL: http://fano.gov.ru/common/upload/library/2014/12/main/kriterii_journals.pdf.

[3] Когаловский М.Р. Метаданные в компьютерных системах// Программирование. 2013. Т. 39, № 4. С. 28–46.

[4] Gantz J., Reinsel D. The Digital Universe in 2020: big data, bigger digital shadows, and biggest growth in the Far East. IDC Digital Universe Study, December, 2012. URL: http://www.emc.com/leadership/digitaluniverse/iview/index. htm.

[5] 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, Вып. 1–2. С. 6–31.

[6] Мбого И.А., Прокудин Д.Е., Чугунов А.В. Комплексная интеграция цифровых коллекций в информационное пространство научных исследований // Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Материалы XVII Всероссийской объединенной конференции «Интернет и современное общество» IMS-2014, Санкт-Петербург, 19–20 ноября 2014 г. С. 48–53.

[7] 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.: Издательство Московского университета, 2014. 262 с.

[8] Елизаров А.М., Липачёв Е.К., Хайдаров Ш.М. Семантический анализ больших коллекций научных документов // Семантические модели и технологии. XIV Международная конференция по компьютерной и когнитивной лингвистике TEL'2016. Казань: АН Республики Татарстан. 2016. С. 5–8.

[9] Елизаров А.М., Липачёв Е.К., Хайдаров Ш.М. Автоматизированная система структурной и семантической обработки физико-математического контента // Ученые записки Института социальных и гуманитарных знаний.

Материалы VIII Международной научно-практической конференции «Электронная Казань 2016» (ИКТ в образовании: технологические, методические и организационные аспекты их использования). Казань: Юниверсум, 2016. № 1 (15). С. 190–197.
[10] Елизаров А.М., Зуев Д.С., Липачёв Е.К. Информационные системы автоматизации цикла подготовки электронных научных журналов (Electronic Scientific Journal Management Systems). Научно-техническая информация. Серия 1, 2014, № 3, С. 31–38 (англ. пер.: Scientific and Technical Information Processing. 2014, V. 41, No 1. P. 66–72).
[11] Stranack K. Getting found, staying found, increasing impact enhancing readership and preserving content for OJS journals. Public Knowledge Project, 2006. URL: https://pkp.sfu.ca/files/GettingFoundStayingFound.pdf.
[12] CrossRef [Электронный ресурс] / Public Knowledge Project [сайт] URL: https://pkp.sfu.ca/crossref/ (дата обращения 06.06.2016).

Automated methods of metadata extraction from scientific publications for bibliographic databases

A.N. Gerasimov, A.M. Elizarov, E.K. Lipachev, S.M. Khaydarov
Kazan Federal University,

An algorithm for automatic extraction of bibliographic data from a one-dimensional array of publications and the formation of metadata blocks for export to international information-analytical system.

Keywords: publishing systems, digital scientific journal, the integration of electronic resources, databases, scientific citation, metadata extraction keywords separated by semicolons.