Проект «Научная электронная библиотека eLibrary.ru» и российские электронные журналы: новый этап развития

© Глухов В. А., Елизаров А. М.

ИНИОН РАН, ООО «НЭБ»,
НИИ математики и механики им. Н. Г. Чеботарева
Казанского государственного университета
vglukhov@inion.ru, elizarov@ksu.ru

Аннотация

Среди многих проектов в области электронных библиотек, которые были реализованы в России за последние годы, проект Научная электронная библиотека (НЭБ) eLibrary.ru занимает особое место. По объему электронного фонда и количеству читателей уже сегодня НЭБ может сравниться с крупнейшими мировыми электронными библиотеками, содержащими научную литературу. Пользователями библиотеки являются практически все ведущие научные организации и университеты страны.

В тесном сотрудничестве с Казанским государственным университетом Научная электронная библиотека уделяет большое внимание развитию российских электронных публикаций, решая, прежде всего, технологические, программные и юридические проблемы. В рамках нескольких грантов Российского фонда фундаментальных исследований (РФФИ) и Министерства образования и науки было разработано несколько программно-технологических комплексов подготовки электронных журналов, с помощью которых только в 2005 году а НЭБ были размещены более 200 российских научных журналов..

1 Электронные научные журналы

Научная электронная библиотека (НЭБ) eLibrary.ru была создана в 1998 году в результате выполнения одного из проектов РФФИ и впоследствии приобрела характер национального проекта. История формирования и развития НЭБ подробно отражена, например, в [1, 2].

В конце 20-го века развитие информационно-телекоммуникационных технологий и средств представления информации привело к тому, что появились электронные версии научных журналов, чтение которых сначала было возможно лишь на отдельных компьютерах и в локальных сетях, а вскоре и через Интернет. Западные издатели стали предлагать своим подписчикам электронные версии журналов как дополнительную услугу. Впоследствии электронные версии стали реализовываться независимо от их печатных аналогов.

В 1996 году в мире существовало всего 250 электронных журналов, однако уже спустя два года их количество увеличилось до 2 тысяч. Сегодня издается, по разным оценкам, от 20 до 30 тысяч электронных научных журналов. Все крупнейшие западные издательства публикуют свои журналы в электронном виде. Появилось значительное количество электронных журналов открытого доступа (см., например, http://www.doaj.org/). Подавляющее большинство электронных журналов доступно через Интернет и объединено в крупные базы данных.

Создание и широкое распространение электронных журналов стало возможным благодаря развитию технологий электронного книгоиздания, специализированных форматов (в основном на основе SGML и XML), средств телекоммуникаций и Интернет, программных средств обработки данных. Однако появление электронных журналов обуславливалось не только этими факторами. Сюда можно добавить также высокую стоимость подписки на печатные версии журналов и ограничение в распространении традиционных изданий. Например, из-за высокой стоимости печатных изданий многие журналы и книги не приобретаются российскими библиотечными учреждениями или доступны в регионах в одной – двух библиотеках; 75% объема рынка российских печатных СМИ, по оценкам Гильдии издателей периодической печати, приходится на Москву. Преобладающий объем столичного рынка обусловлен также тем, что издания не могут пробиться в регионы из-за отсутствия системы распространения печатной периодики. Для электронных журналов нет таких ограничений, хотя система подписки на электронные издания в стране почти полностью отсутствует.

Труды 7ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2006, Суздаль, Россия, 2006.

2 Система электронного книгоиздания в России

В области электронного книгоиздания Россия значительно отстает от других стран. В 2005 году нами была проведена экспертиза почти тысячи журналов, входящих в список Высшей аттестационной комиссии (ВАК) РФ. Из них в Интернете так или иначе присутствовало около 300 журналов. Однако детальное изучение сайтов показало, что большинство журналов представлено только в виде оглавлений с аннотациями. Часть журналов представлена 1 – 2 годами изданий, публикация которых в электронном виде была осуществлена несколько лет назад. Количество «действующих» электронных журналов из списка ВАК (т. е. тех, которые были опубликованы в 2004 – 2005 гг.) не превышало 100 наименований. Практически на всех сайтах отсутствует возможность поиска по авторам, названиям статей, ключевым словам, аннотациям и особенно – по полным текстам статей. Тексты представлены в разных форматах – html, PDF, DjVu, Word и т. д. Неразвитость российских электронных научных журналов приводит к тому, что ученые стали отдавать предпочтение публикации своих работ в западных журналах, что в свою очередь сказывается на авторитете наших изданий и ведет к их упадку. С этой точки зрения очень важны развитие технологий электронного книгоиздания в России, появление своих программно-технологических комплексов подготовки электронных изданий. Среди причин недостаточного развития электронного книгоиздания можно назвать следующие.

- Отсутствуют форматы представления электронных изданий, предполагающие детальное разбиение элементов полнотекстовых журналов, которые могут быть использованы, в том числе, для загрузки журналов в базы данных. Форматы записей, давно и успешно применяемые в библиотечном деле, такие, как US Marc, Unimarc или Rusmarc, предназначены лишь для описания печатных источников (книг, журналов в целом, статей) на библиографическом уровне. Однако в случае описания полных текстов журналов или отдельных публикаций эти форматы не могут быть использованы из-за отсутствия необходимых спецификаций. Прежде всего, поскольку библиотечная каталогизация направлена только на описание печатных источников, то эти форматы (при всей их детализации) зачастую не включают описания таких элементов, как, например, подробные сведения об индивидуальных авторах (ученое звание, место работы, почтовый и электронный адрес). Ни один из библиотечных форматов не поддерживает описаний полного текста статьи из журнала, главы из книги или текста книги целиком. Полностью отсутствует описание (тем более детализированное) пристатейных или прикнижных списков использованной литературы, что весьма важно при построении индексов научного цитирования. Точно так же не могут применяться описания, построенные на языках метаописания, подобных Dublin Core (DC) (см. http://purl.oclc.org/dc/), используемых в основном для библиографического описания Интернет-страниц. Поэтому требуется разработка форматов, предназначенных специально для описания полнотекстовых электронных изданий.
- Отсутствует специализированное программное обеспечение, позволяющее производить углубленную обработку электронных версий журналов с целью их последующей загрузки в базы данных. Известно, что все научные журналы, издаваемые в стране, сейчас создаются с использованием компьютерной техники. Применяется достаточно много издательских пакетов программ (PageMaker, Word и др.). Однако почти все они предназначены лишь для набора текста, оформления и создания макетов изданий, с которых производится типографское тиражирование. Использование таких макетов для загрузки электронных изданий в базы данных не предусматривается и не может быть выполнено. Необходима разработка программно-технологических комплексов, предназначенных для структурирования (разделения на поля) электронных версий печатных изданий.
- Издательства и редакции не очень охотно идут на публикацию электронных изданий. Во многом это связано с опасением утратить подписку на печатный вариант издания. Многие проблемы электронной публикации журналов не имеют достаточной юридической базы. Необходимо предложить владельцам журналов разумные схемы взаимодействия с электронными библиотеками, включая подготовку пакетов юридических документов по прямому лицензированию. Кроме того, необходимо подготовить юридическую документацию, регулирующую проблемы авторского права применительно к электронным публикациям – на уровне «автор – редакция» или «автор – издательство».

По проекту Министерства образования и науки «Создание электронных версий российских научных журналов» (шифр РИ-13.0/006) в 2005 году Казанским государственным университетом (КГУ) и Научной электронной библиотекой были разработаны программы и технологии обработки электронных версий печатных журналов, созданы форматы данных, подготовлен пакет юридических документов, регулирующих отношения авторов, издателей и электронных библиотек. При этом решались задачи обработки как печатных изданий, не имеющих «представительства» в Интернет, так и электронных журналов.

Одной из целей исследований, проведенных в рамках совместных проектов РФФИ (проекты 02-07-90230 и 03-07-90252) и Российского гуманитарного научного фонда (проект 03-03-12007) в 2003 – 2005 гг., была разработка полнофункциональной программной среды электронного научного журнала по математике, позволяющей автоматизировать ряд процессов, стандартных для научного издания.

С точки зрения обработки информации процесс электронного издания представляет собой

XML/MathML-поток (http://www.dessci.com/en/products/mathflow/). На входе этого потока находятся материалы, представленные авторами в LaTeX-формате с использованием стандартных для журнала стилевых правил (например, http://ljm.ksu.ru/submission.html), на выходе – очередной том журнала с возможностью выбора для читателя подходящего формата представления данных (pdf, djvu, html и т. д). Данная схема реализована в электронном научном журнале «Lobachevskii Journal of Mathematics» (LJM) (электронный адрес журнала – http:/ljm.ksu.ru, ISSN – 1818-9962). В настоящее время журнал представляет читателям статьи в форматах pdf, ps, dvi, а последние тома содержат также статьи в формате MathML.

В журнале LJM с использованием ряда технологических приемов организован XML/MathML-поток получения, преобразования и хранения данных. Один из разрезов потока представляет собой преобразование представленных авторами материалов (LaTeX-документов и данных используемых форм) в XML-файлы со вставками MathML. Формы для авторов содержат сведения об авторе и ключевые слова, отражающие содержание публикации. XSL-модули отвечают за преобразование информации, содержащейся в XML-хранилище, в представление, выбранное читателем (например, предназначенное для печати). Кроме того, рассматривались вопросы конвертации в MathML документов, подготовленных с помощью имеющихся стандартных технологий (LaTeX, Mathematica, Maple, Word). Язык метаописания Dublin Core на основе Resource Description Framework (RDF) (см. http://www.w3.org/TR/2002/WD-rdf-syntax-grammar-20021108) был использован для включения метаданных в xml-файлы. Это позволило обеспечить доступ поисковых машин к библиографическим данным статей, опубликованных в журнале. Подробное описание этих результатов можно найти, например, в [3]. В настоящее время эти исследования проводятся при финансовой поддержке РФФИ (проект 06-07-89132).

3 Форматы для электронных журналов и книг

Как уже отмечалось, российские форматы представления электронных изданий, предполагающие детальное разбиение элементов изданий, которые могут быть использованы для загрузки полнотекстовых журналов и книг в базы данных, в настоящее время отсутствуют. Издатели электронных журналов сейчас применяют, как правило, html-разметку выпусков своих изданий, не следуя при этом никаким общепринятым правилам, поскольку их просто не существует. Кроме того, такая разметка не может быть использована для загрузки в базы данных.

Для структурирования макетов печатных изданий в рамках проекта «Научная электронная библиотека eLibrary.ru» была разработана программа, в основу которой положен принцип выделения элементов текста и присвоения им меток полей собственного XML-формата, названного Sarcticle.

Отличительными особенностями этого формата являются: вложенность полей, возможности описания любого количества информации одним файлом, проверки правильности составления файлов описаний на стороне издательств, использования файлов описаний для наполнения собственных сайтов издательств и совместимости с другими форматами обмена метаданными, основанными на XML. Основные блоки формата – информация о журнале, о выпуске, о статье (основная информация файла). Большинство полей может дублироваться на нескольких языках с целью более удобного представления для разных пользователей конечной информации в электронной библиотеке.

Основные разделы формата:

- раздел описания журнала в целом, куда входят сведения о названии журнала, издателе, ISSN, обобщенной структуре издания (том – номер – часть – спецвыпуск), а также поля, позволяющие описать отдельный выпуск журнала;
- сведения о статье из выпуска журнала, куда входят описание индивидуальных и/или коллективных авторов статьи с подробной информацией о них, название статьи, ключевые слова, реферат (аннотация), полный текст статьи без списка литературы, наиболее распространенные коды классификаторов (УДК, ББК, ГРНТИ, DOI для электронных изданий и др.), а также подраздел, описывающий пристатейные списки литературы; при этом каждая позиция в списке литературы (или сноске) разбита на отдельные поля и подполя – например, автор(ы) работы, название, источник, год издания и т. д.;
- раздел тематических рубрик журнала, куда входит описание подразделов выпуска журнала.

Формат исполнен в двух видах – в DTD и в MS Schema. Набор тагов формата не зависит от выбора видов описания XML. Порядок следования тагов важен. Все таги имеют закрывающий таг. Регистр тагов должен соблюдаться: используются как строчные, так и прописные буквы в названиях тагов. Все спецсимволы при использовании формата требуется заменить на предопределенные сущности.

Технически возможно в одном файле описать любое количество журналов, но с точки зрения удобства хранения и заполнения предпочтительна ситуация «один файл XML – один выпуск журнала».

Возможные способы создания документов XML в формате Sarcticle могут включать использование:

- специализированных программных средств создания документов XML, конформных формату Sarcticle;
- любого XML-ориентированного текстового редактора, например, MS XML Notepad;
- любого текстового редактора.

Имеются дополнительные описания элементов формата (или «справочники»):

- «arcticle types» – список кодов типов статей для атрибута arttype;

- «language codes» – список кодов языков для атрибута fieldlang;
- «country codes» – список кодов стран для атрибута jcountry;
- «symbols.html» (в HTML) – список всех сущностей, заменяющих специальные символы;
- «dateUni format.txt» – описание формата поля dateUni.

4 Программное обеспечение разметки электронных журналов и загрузки в базу данных электронной библиотеки

Все научные журналы сейчас готовятся к изданию с использованием компьютерной техники. Применяется большой перечень издательских пакетов программ. К сожалению, за редким исключением эти программы предназначены лишь для набора текста, оформления и создания макетов изданий, с которых производится типографское тиражирование. Использование макетов для загрузки электронных изданий в базы данных не может быть выполнено. Необходимо было разработать специальные программы структурирования (разметки) текстов электронных версий журналов для создания файлов, предназначенных для загрузки в базы данных. Попытки создания специальных программ обработки макетов изданий предпринимались неоднократно, однако ни одна из них не увенчалась успехом. Проблема в том, что корректная работа подобных программ с макетами изданий возможна лишь в том случае, если сам макет сделан при тщательном соблюдении правил, что практически неосуществимо. Отчасти этим условиям соответствуют издания, подготовленные с использованием LaTex, однако и в этих случаях требуется создание специализированных конверторов. Но, самое главное, что и в этом случае многие значимые элементы издания не могут быть программно выделены из текста и структурированы.

Для решения проблемы были разработаны три версии программного обеспечения для разметки макетов электронных изданий, в основу которого был положен принцип выделения элементов текста макетов изданий и присвоения им меток полей описанного выше XML-формата Sarcticle.

Входными данными для программы являются электронные документы в формате MS Word (rtf и doc), Adobe Acrobat (pdf)[1], текстовые файлы (html и txt), содержащие полные версии журналов, сборники статей или иную библиографическую информацию. Кроме того, возможна обработка Интернет-страниц журналов.

Разметка такого документа производится последовательным выделением в его тексте элементов библиографического описания. Каждый выделенный элемент связывается с соответствующим полем из набора полей, определенных форматом. На этапе такого выделения и связывания элементов формируется визуальный образ будущего xml-документа, представляющий собой пакет древовидных иерархических структур.

Существует два типа таких структур: для отображения библиографических данных журнала и для отображения библиографических данных статьи из журнала. Каждая такая структура формируется как самостоятельный визуальный объект и размещается на отдельной странице. Программа производит верификацию и конвертацию данных с формированием выходного xml-документа отдельного выпуска (номера) журнала.

Был разработан специальный модуль программного обеспечения, позволяющий производить автоматическое структурирование (разбор по полям формата) пристатейных списков литературы и сносок. Разработка такого модуля необходима для подготовки библиографических материалов, включаемых в различные индексы научного цитирования.

Модуль поддерживает работу с пристатейными списками литературы, которые подготовлены в соответствии с ГОСТ 7.1-84 «Библиографическое описание документа». Работа модуля заключается в автоматическом выделении из библиографического описания книги, статьи из журнала или сборника, патента, авторского свидетельства, на которые ссылаются авторы оригинальных статей. Производится выделение следующих полей: автор произведения, название, источник (название журнала или сборника), место издания и издательство, год издания, том, номер, часть, страницы, а также URL-адрес электронной версии источника.

Программное обеспечение разметки электронных журналов прошло успешное тестирование в ряде редакций научных журналов. Однако необходимо было создание новой версии программы, позволяющей работать с большим количеством исходных форматов (html, PageMaker, PDF, Word и другими) и максимально автоматизирующей процесс структуризации текста макетов. Модифицированная версия программы, созданная в рамках выполнения проекта, позволила расширить перечень обрабатываемых форматов исходных макетов изданий. Кроме того, добавлены два новых специальных модуля для программного распознавания группы авторов с их описаниями (место работы, почтовый и электронный адреса и т. д.), ключевых слов, а также списков пристатейной литературы.

Одной из основных проблем, которая решалась в ходе разработки программы разметки текстов электронных журналов, были обработка и отображение математических и химических формул, диакритических, математических и других специализированных знаков и символов. Новая версия программы позволяет обрабатывать такие макеты и создавать XML-файлы как в формате Ascii, так и в формате Unicode (UTF-8 и UTF-16).

В результате разработки и использования программы время обработки усредненного выпуска журнала (115 страниц, 3 автора каждой статьи, 17 –

[1] Требуется предварительная обработка исходных файлов в этом формате

18 статей и 15 – 20 пристатейных ссылок) составляет 2 – 3 часа рабочего времени.

Загрузка данных в формате Sarcticle-XML в базу данных реализована с использованием технологии Windows Script Host (WSH). Создан универсальный загрузчик на VBScript, осуществляющий парсинг данных и записывающий результат в базу данных. При этом исходные тексты статей в форматах pdf и/или html записываются на файловый сервер. Апробация программ загрузки данных в базу электронной библиотеки выполнялась на 60 выпусках журналов издательства Института научной информации по общественным наукам.

В результате загрузки журналов в формате Sarcticle пользователи электронной библиотеки могут проводить поиск статей из журналов по следующим параметрам: авторы, названия статей, аннотации (рефераты), ключевые слова, слова из полных текстов статей, библиографические описания источников из пристатейных списков литературы.

Заключение

По разработанной технологии, описанной выше, в НЭБ были размещены более 200 научных журналов. Среди них такие авторитетные издания, как «Успехи физических наук», «Успехи химии», «Биология моря», «Ученые записки Казанского университета», «Известия вузов. Авиационная техника», «Известия вузов. Математика», «Известия вузов. Радиофизика», «Казанский медицинский журнал», «Вестник Казанского государственного технического университета», «Вестник Дальневосточного отделения РАН», «Дальневосточный математический журнал», «Тихоокеанская геология» и ряд других. Большинство из журналов до 2005 года не было представлено в Интернете.

Некоторые разработки в рамках совместных проектов НЭБ и КГУ были положены в основу создаваемого Научной электронной библиотекой Российского индекса научного цитирования (см. http://www.elibrary.ru/projects/citation/ proposal.doc).

Литература

[1] Глухов В.А., Новиков В.Д., Петров А.Н. Научная электронная библиотека: итоги и перспективы // Вестник Национального комитета «Интеллектуальные ресурсы России». – Москва, 2004, № 2, С. 9-14. – URL: http://www.elibrary.ru/item.asp?id=8826534.

[2] Глухов В.А. Проект «Научная электронная библиотека eLibrary.Ru» и перспективы развития электронного книгоиздания в России // Educational Technology & Society, 2005, Т. 8, № 1, С. 191-197. – URL: http://www.elibrary.ru/item.asp?id=8370428.

[3] Елизаров А. М., Липачев Е. К., Малахальцев М. А. Технологии Semantic Web в практике работы электронного журнала по математике//Настоящий сборник.

Project Elibrary.ru and Russian Electronic Journals: New Stage of Development

Glukhov V. A., Elizarov A. M.

In this paper we the results on development of the system of electronic publishing of books in Russia, which were obtained in 2003 – 2005 in the framework of the project Elibrary.ru.