Технологии Semantic Web в практике работы электронного журнала по математике

© Елизаров А. М., Липачёв Е. К., Малахальцев М. А.

НИИ математики и механики им. Н. Г. Чеботарева
Казанского государственного университета
elizarov@ksu.ru

Аннотация

Изложены подходы к автоматизации редакционной обработки статей электронного научного математического журнала на основе XML, RDF и других технологий Semantic Web. В частности, решены вопросы генерации метаданных, организации хранения и поиска данных, автоматической конвертации из формата tex в форматы pdf, ps, mathml. Эти подходы реализованы в электронном журнале «Lobachevskii Journal of Mathematics» (http://ljm.ksu.ru).

1 Электронные научные журналы

Перевод научных знаний в электронную форму приобретает все большее значение по мере освоения компьютерных технологий научным сообществом. Изменяются сложившиеся формы обмена научной информацией: проводятся виртуальные конференции, неуклонно растет число электронных научных архивов и баз данных. Эволюция форм научного обмена, вызванная развитием сетевой инфраструктуры, привела, в частности, к созданию электронных научных журналов.

Необходимым условием функционирования электронного журнала является признание научным сообществом равноценности электронной и традиционной (в «бумажном» научном журнале) публикаций. Это накладывает на редакции электронных журналов ту же ответственность, что и в любом традиционном научном журнале, в частности, по организации независимого научного рецензирования.

Модель работы редколлегии электронного научного журнала, сходная по многим общим чертам с работой традиционного журнала, имеет ряд существенных особенностей – достаточно указать на качественно иной уровень оперативности на всех этапах редакционной работы с электронной публикацией. Необходимость оперативного рассмотрения статей редколлегией электронного научного журнала неизбежно приводит к отказу от привычных форм работы и переходу к электронным методам документооборота. Автоматизация даже части работы редколлегии электронного журнала требует решения множества задач и привлечения современных сетевых технологий.

В настоящей работе описан вариант автоматизации редакционной обработки статей, разработанный для электронного математического научного журнала «Lobachevskii Journal of Mathematics» (LJM) (электронный адрес журнала – http://ljm.ksu.ru; ISSN журнала – 1818-9962).

LJM – один из первых отечественных научных электронных журналов – издается с 1998 года (дата регистрации в Министерстве по делам печати и информации – 2 августа 1996 года). За это время вышло 20 томов журнала. Статьи российских авторов занимают примерно половину объема журнала. В журнале публикуются также работы авторов из Армении, Индии, Испании, Канады, Китая, Марокко, Норвегии, Сербии, США, Финляндии, Японии. Редколлегию журнала возглавляет крупнейший российский математик, академик РАН С. П. Новиков. В составе редколлегии – ведущие специалисты в большинстве областей современной математики. LJM представлен на сервере Европейского математического общества, статьи реферируются в журналах Mathematical Review, Zentralblatt fur Mathematic, информация о журнале регулярно появляется в Notices of American Mathematical Society. В настоящее время журнал включен в базы данных Science Direct издательства Elsiever (http://www.elsiever.com) и Научной электронной библиотеки E-library (http://elibrary.ru).

Особенностью журнала является разнообразие его тематики: в журнале представлены работы по алгебре, геометрии и то-

пологии, комплексному анализу, функциональному анализу, теории вероятностей и математической статистике, оптимальному управлению, теории алгоритмов. Это создает дополнительные сложности в процессе автоматизации документооборота редколлегии электронного журнала.

В силу указанных обстоятельств руководство журнала особое внимание уделяет внедрению современных информационных технологий. Разработан и внедряется автоматизированный комплекс редакционной обработки статей на основе технологий Semantic Web. Он позволяет производить регистрацию статьи и ее предварительную обработку, в том числе преобразование в форматы представления. Наряду со стандартными наборами форматов представления математических статей (dvi, pdf, ps) журнал представляет статьи в формате MathML (впервые в мире среди электронных журналов). Это потребовало адаптации имеющихся программных средств конвертации TeX-файлов в MathML-файлы и встраивания их в автоматизированную систему публикации статей.

LJM функционирует в системе международных электронных научных журналов, издаваемых Казанским университетом («Journal of Formal, Computational and Cognitive Linguistics», LJM, «Magnetic Resonance in Solids», «Информационные технологии и телерадиокоммуникации»). Переход на новые технологии позволил вывести работу этих журналов на совершенно иной уровень, приближающийся к современным информационным стандартам, и существенно повысить эффективность их использования. Отметим, что в работе журнала LJM применены подходы, которые в определенной степени отличаются от принятых. Сравнительный анализ используемых технологий позволит определить оптимальную стратегию развития электронных журналов.

Разработанные нами технологии, апробированы в «Lobachevskii Journal of Mathematics» и предлагаются в качестве возможного варианта автоматизации работы редколлегий электронных журналов, прежде всего, математических.

2 Semantic Web в электронном математическом журнале

Semantic Web разрабатывается консорциумом W3C (http://www.w3c.org) как перспективная «машиноориентированная» технология, способная заменить традиционные Internet-технологии, требующие непосредственного участия человека в большинстве операций по обработке данных (см., например, [1]). Текущие проблемы, связанные с развитием Semantic Web, постоянно обсуждаются специалистами (см., например, материалы конференции WWW2002 Workshop on Real World RDF and Semantic Web Applications – http://www.cs.rutgers. edu/~shklar/www11/). Поддержка технологий Semantic Web реализуется в ряде крупных исследовательских проектов (например, http://www.ilrt. bris.ac.uk/projects/semantic_web, http://www.ilrt.bris. ac.uk/people/cmdjb/), направленных на обеспечение семантической интероперабельности в Web.

Диаграмма «SemanticWeb layer cake», предложенная Бернерсом-Ли (http://www.w3.org/2000/ Talks/1206-xml2k-tbl/slide10-0.html; см. также [2]), дает представление о многослойной архитектуре Semantic Web, включающей, в частности, Extensible Markup Language (XML), Resource Description Framework (RDF), RDF Schema и Web Ontology Language (OWL). Структурирование информации производится на основе XML, а RDF используется для создания расширенной системы метаданных на основе иерархических моделей.

Выбор технологий Semantic Web для организации работы математического электронного журнала объясняется наличием в них инструментов, позволяющих записывать данные, учитывая структурную и семантическую составляющие информации. Наиболее полно современное представление о методологии применения электронных технологий к формированию и хранению специализированной научной информации отражено в рекомендациях консорциума W3C (www.w3.org), а также в разработках корпорации Wolfram Research [3].

Программные решения, применяемые в настоящее время при подготовке научных изданий, основаны на html и, в значительной степени, предполагают участие человека в обработке информации. Поскольку тэговая разметка языка html позволяет структурировать текст только в части отображения документа, говорят о слабострук-

турированном представлении информации. Подготовка математических текстов производится в TeX-формате (обязательное требование большинства математических журналов). Это текстовый формат с тэговой разметкой, обеспечивающей форматирование документа и включение математических формул (http://ctan.org). TeX-формат также является слабоструктурированным. Автоматизация процедур обработки слабоструктурированной информации затруднительна и не всегда эффективна. В частности, поиск информации, хранящейся в слабоструктурированных форматах нельзя полностью автоматизировать из-за сложности процедуры извлечения данных.

Отдельной задачей, стоящей перед научными изданиями, является обеспечение возможности извлечения метаинформации («информации об информации») поисковыми системами, большая часть которых для индексирования применяет программы-роботы, использующие метаописание ресурсов сети. Метаданные содержат обобщенную информацию о структуре и содержании информационного источника (автор, дата, источник, ключевые слова, предметная область и т. д.). В формате html описание метаданных возможно только через метатэги. Для наиболее унифицированного описания и каталогизации ресурсов в сети создаются специальные метаязыки, наиболее распространенным из них является Dublin Core (DC). Однако слабоструктурированная информация не дает возможности автоматизировать подготовку метаданных. Более того, для составления блока метаданных каждого документа требуется участие квалифицированного специалиста.

Основная проблема при создании, хранении и отображении электронных публикаций по математике касается представления математических формул. Наиболее распространенное на данный момент решение – представление формул в виде графических файлов – неудовлетворительно с точки зрения структурной обработки математических текстов [4].

Для решения указанных проблем можно применить стандарты и рекомендации, разработанные консорциумом W3C в рамках проекта Semantic Web (см. также [5]). В частности, в 1999 году консорциумом W3C (http://www.w3.org) была начата разработка языка математической разметки – Mathematical Markup Language (MathML). Этот язык представляет собой подмножество языка разметки XML (см. [6] – [9]), который позволяет решать задачи программной обработки документов (в частности, задачу поиска) на новом технологическом уровне. Поэтому использование XML, в частности, MathML, изменяет принципы организации и управления электронными публикациями по математике.

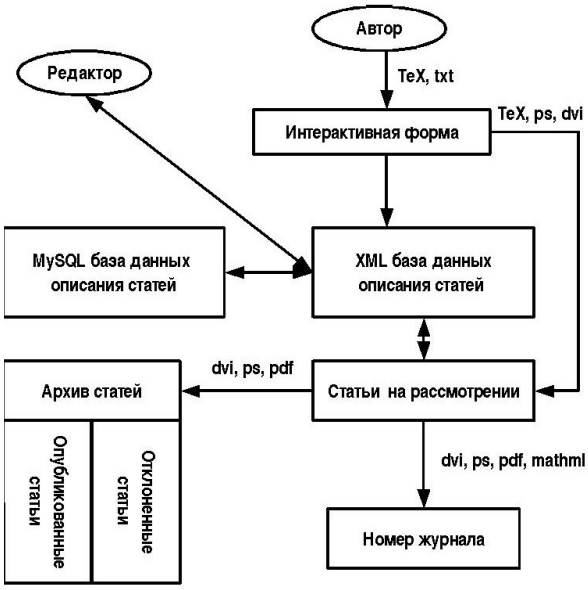
В настоящее время язык MathML становится стандартом представления математической информации в электронной форме в силу следующих причин: технология обработки данных на основе языка MathML реализует одну из основных тенденций современной информатики – разделение разметки и данных, поэтому она представляет широкие возможности многоуровневого структурирования данных и расширенного поиска; имеется возможность создания программного обеспечения, использующего технологию MathML; созданы и продолжают совершенствоваться программные средства, позволяющие конвертировать в MathML документы, подготовленные с помощью имеющихся стандартных технологий (LaTeX, Mathematica, Maple, MS Word). Из наиболее распространенных инструментов отметим редактор WebEq (www.dessci.-com/en/products/ webeq), конверторы TeX в MathML: TtM (http:// hutchinson.belmont.-ma.us/tth/mml/), TeX4ht (входит в стандартный пакет MikTeX), а также конвертор XML в TeX (MikTeX). MathML поддерживается основными просмотрщиками: Internet Explorer (при установке соответствующего модуля [10, 11]), Mozilla, Amaya; технология MathML поддерживается системами Maple® и MathCAD 2001, а компания Wolfram Research предложила собственную концепцию использования технологии MathML [12], которая реализована в пакете Mathematica® [3, 12], в частности, в этом пакете предусмотрено сохранение документов в формате MathML.

3 Автоматизация работы электронного журнала

Работа по автоматизации электронного журнала велась на основе CASE-моделирования. Была создана модель электронного математического хранилища, состоящая из комплекса UML-диаграмм.

Разработаны методы автоматической обработки электронной математической информации на основе архитектуры Semantic Web. Семантическая разметка Web-страниц выполняется на основе стандарта XML, при этом для разметки математической части текста использован MathML. Разработана иерархическая модель метадан-

ных для ресурсов электронного математического журнала в соответствии со стандартом RDF. Разработаны методы использования технологии XSLT для преобразования математических текстов и служебной информации, представленной в XML-формате, в частности, MathML. Спроектирована архитектура таблиц MySQL базы данных электронного математического журнала. Согласно разработанной схеме, в XML-файле накапливается информация, поступающая в результате пользовательского запроса, далее применяется XSLT-преобразование и результат перенаправляется в MySQL-базу. Это технологическое решение позволило существенно снизить нагрузку (количество запросов) на MySQL-сервер, поскольку каждый запрос работает с отдельным XML-файлом, без постоянного подключения к базе.



На диаграмме представлена принципиальная схема процесса редакционной обработки статьи. Каждому прямоугольнику отвечает программный модуль, выполняющий соответствующую процедуру. На стрелках указаны форматы, участвующие в операциях обмена данными.

Указанные подходы были реализованы в виде программного комплекса для обработки и управления потоками данных в электронном журнале на основе XML/XSLT и MathML-технологий. Одной из составляющих этого комплекса является система PHP-скриптов, которая включает в себя, в частности, сервисы автоматизированного представления рукописей статей в журнал и автоматизированного прохождения рукописи. Ряд задач был решен с помощью скриптов на языках AWK и Perl. Для генерации и обработки запросов используется язык RDQL [13]. Создано программное обеспечение, осуществляющее автоматический перевод в RDF-формат метаданных, поступающих из интерактивной формы, предоставленной пользователю. Преобразование XML/RDF-информации в html-формат производится с помощью скриптов (Java, PHP). Осуществлена программная реализация управления потоками данных с Web-сайта в MySQL-базу с использованием XML-файлов в роли буфера.

Сервис автоматизированного представления рукописей включает в себя регистрацию авторов и внесение их данных в MySQL-базу; сбор метаданных из формы, заполняемой авторами публикаций; компиляцию TeX-файлов в форматы dvi, ps, pdf для редакторской работы; внесение основных данных о рукописи (название, AMS-классификация, ключевые слова, дата подачи рукописи) в MySQL-базу; извещение секретаря журнала о поступившей рукописи.

Сервис автоматизированного прохождения рукописей позволяет получить информацию о статусе статьи (наличие отзывов с датами представления), дает возможность технического редактирования рукописи, генерирует файлы в форматах dvi, ps, pdf, mathml, а также генерирует html-файлы с информацией о статьях и о томе журнала. Также организовано автоматическое генерирование метаданных по технологиям DublinCore и RDF.

Разработанное программное обеспечение позволило осуществить перевод в формат MathML полных текстов ранее опубликованных статей. В настоящее время статьи хранятся как в формате MathML, так и в стандартных форматах (dvi, ps, pdf). Кроме того, на сайте журнала LJM размещены аннотации статей в формате MathML.

Работа с сайтом журнала предполагает, что клиент выполнит необходимые настройки своего браузера. Прежде всего, требуется обеспечить поддержку MathML. На сайте журнала (http://ljm.ksu.ru) имеются инструкция по настройке наиболее распространенных браузеров и список узлов с необходимым программным обеспечением.

В связи с тенденцией перехода на открытые Unix-подобные операционные системы разработанное программное обеспечение адаптировано к работе под управлением ОС Linux и использует средства Unix-ориентированных систем.

В заключение отметим, что разработанные программные средства внедрены в работу электронного журнала LJM. В настоящее время все публикуемые в журнале

статьи снабжаются комплексом метаданных, включающим смешанный набор метатэгов. Этот набор состоит из традиционного набора метаданных и метаданных в формате DC. Это позволяет индексировать страницы журнала как поисковым системам, роботы которых распознают только традиционный формат метаданных, так и поисковым системам, рассчитанным на метаданные в формате DC.

Работа поддержана РФФИ, проект 06-07-89132.

Литература

1] Berners-Lee T. Semantic Web Road Map – http:// www.w3.org/DesignIssues/Semantic.html).

2] Passin T. M. Explorer's Guide to the Semantic Web//Manning Publication Company, 2004 – 280 с.

3] Wolfram Research Contributes Central Ideas to Web Math Standard. – http://www.wolfram.com/ ews/archive/mathml.html.

4] Митюнин В. А. Обзор средств публикации и просмотра математических документов в сети Интернет – http://math-mag.spbu. ru/article/4/.

5] Елизаров А. М., Липачев Е. К., Малахальцев М. А. Электронные журналы по математике и рекомендации консорциума W3// Труды. Всерос. науч. конф. «Научный сервис в сети Интернет», г. Новороссийск, 22 – 27 сент. 2003 г. – М.: Изд-во Московского ун-та, 2003. – С. 75-76 (http://agora.guru.ru/).

6] MathML 1.01 – http://www.w3.org/TR/REC-MathML/.

7] MathML 2.0 – http://www.w3.org/TR/MathML2/.

8] Елизаров А. М., Липачев Е. К., Малахальцев М. А. Основы MathML. Представление математических текстов в Internet. Практическое руководство. – Казань: Изд-во Казан. матем. об-ва, 2004. – 60 с.

9] Sandhu P. The MathML Handbook. – Charles River Media, 2003 (http://www. tephenwolframcom/ews/).

10] MathPlayer Display MathML http://www.dessi. com/ n/product/ mathplayer.

11] TechExplorer – http://www-3.ibm.com/software/ network/techexplorer/.

12] Wolfram Stephen. Mathematical Notation: Past and Future. – http://www.stephenwolfram.com. /publications/talks/mathml/.

13] RDQL - A Query Language for RDF/ http://www. w3.org/Submission/2004/SUBM-RDQL-20040109/.

Semantic Web Technology for Electronic Mathematical Journal

Elizarov A. M., Lipachev E. K., Malakhaltsev M. A.

In this paper we describe approach to the automatization of editorial process in an electronic mathematical journal. On the base of XML, RDF and other Semantic Web formats we solve the problems of generating metadata from tex, and automatic convertation from tex to pdf, ps, mathml. The software designed according to this approach, is used in Lobachevskii Journal of Mathematics (http://ljm.ksu.ru).