Сервисы структурирования математического контента и интеграция электронных математических коллекций в научное информационное пространство

© А.М. Елизаров © Д.С. Зуев © Е.К. Липачёв
© М.А. Малахальцев
Институт математики и механики им. Н.И. Лобачевского
Казанского (Приволжского) федерального университета
amelizarov@gmail.com dzuev11@gmail.com lipachev@ksu.ru
mikarm@uniandes.edu.co

Аннотация

Процесс структурирования (разделения на смысловые элементы) электронных версий печатных изданий является необходимым этапом для последующего семантического структурирования и включения электронных коллекций в информационное пространство.

Электронные версии печатных научных журналов представляют собой документы, имеющие структуру, которая отражает логику разделение документа на части. Эта структура сформирована шрифтовым выделением, абзацами, вертикальными и горизонтальными отступами. Автоматическая обработка таких документов с целью отбора структурных компонент (например, выделения авторов статьи или библиографических данных) затруднительна. Как следствие, большинство операций с электронным контентом, в частности, создание связей между объектами электронного хранилища, необходимо выполнять вручную.

В докладе обсуждается подход к автоматизации процесса обработки научных электронных документов и их преобразования в структурированные документы. Акцент сделан на особенностях обработки математических текстов. С помощью сервисов, созданных по предложенной методике, выполнено структурирование достаточно большого по объему электронного хранилища, содержащего выпуски периодического журнала по математике и многотомных трудов конференций.

Работа поддержана РФФИ (проекты № 12-07-00667 и 12-07-97018-р_поволжье)

Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.

1 Введение

Многочисленные научные электронные коллекции, созданные на основе действующих периодических научных журналов, состоят из электронных версий статей, изданных типографским способом. Как следствие, электронные документы в таких коллекциях хранятся в виде отдельных файлов с минимальной метаинформацией и не имеют структуры, позволяющей выполнить автоматизацию по выделению данных.

Как правило, электронные документы не имеют связей внутри коллекции. Аналитическая обработка документов такой коллекции (например, сбор наукометрических данных) представляется крайне затруднительной.

Электронные журналы, публикующие материалы исключительно в электронном виде, в большинстве случаев являются коллекцией электронных документов, созданных теми же программными средствами, ориентированными только на «финишную» печать и, следовательно, имеющими только структуру оформления.

Примерами современных электронных хранилищ с развитыми сервисами являются научная электронная библиотека eLibrary.ru (http://elibrary.ru) и общероссийский математический портал Math-Net.Ru (http://www.mathnet.ru/). Отметим также коллекцию электронного математического журнала Lobachevskii Journal of Mathematics, содержащую сервисы управления электронным контентом, сформированные на основе семантического Веба (см., напр., [1], [2]). Этот журнал, издаваемый с 1998 года, является одним из первых российских электронных научных журналов и включен в базы данных Science Direct (Elsevier) и eLibrary.ru.

На современном этапе развития электронных научных библиотек важное место занимает интеграция созданных научных ресурсов в научное информационное пространство, в котором между объектами электронных коллекций присутствуют семантические связи (см., напр., [3], [4]). Необходимым условием такой интеграции являются семантическое структурирование контента научных электронных

библиотек и создание семантических связей между информационными объектами. Технологии семантического Веба, разрабатываемые консорциумом W3C (www.w3.org), являются технологической платформой, на которой осуществляется интеграция электронных ресурсов в информационное пространство (см., напр., [5], [6]).

2 Проблемы обработки электронных математических ресурсов

Как уже было отмечено, у большинства имеющихся электронных документов, являющихся электронными версиями печатных публикаций, можно обнаружить только структуру, отражающую форматирование (шрифт, выделение). Выполнить преобразование таких документов в структурированный документ можно на основе особенностей форматирования. Процесс такого преобразования можно разделить на несколько последовательных этапов, с которыми связано решение соответствующих задач.

Первая задача – это разделение текстов на категории по общей для них системе форматирования и программным средствам, используемым для научной разметки. Можно считать, что такое разделение уже сделано – журналы, сборники трудов и т. д., как правило, подчинены единообразному для каждого издания стилю оформления. Затруднение может вызвать только система научной нотации – в ряде изданий можно обнаружить, что разные авторы используют отличающиеся технологии разметки. Например, в одном и том же сборнике наряду со статьями, выполненными в TeX-разметке, присутствуют статьи, выполненные в MS Word + MathType.

Следующая задача – создание системы признаков для каждой категории электронных документов, на основании которых из текста выделяются структурные элементы.

Сложной задачей является обработка электронного документа и его трансформация в структурированный документ на основе системы признаков.

Отдельная задача заключается в генерации метаданных и выделении из текста ключевых слов.

Завершающим этапом является создание электронного документа, структурированного по правилам семантического Веба.

3 Технологии структурирования электронных ресурсов

Один из подходов к структурированию макетов печатных изданий в составе электронной коллекции предложен в проекте «Научная электронная библиотека eLibrary.ru». Алгоритм структурирования основан на выделении элементов текста и присвоении им специализированных меток.

Подготовка библиографических материалов, включаемых в индексы научного цитирования, выполняется автоматически с помощью сервиса, производящего структурирование списков литературы и сносок с учетом требований ГОСТ 7.1-84 «Библиографическое описание документа».

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 (см. [7]).

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 тегов формата не зависит от выбора видов описания XML. Порядок следования тегов важен. Все теги имеют закрывающий тег. Регистр тегов должен соблюдаться: используются как строчные, так и прописные буквы в названиях тегов. 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.

В случае электронной коллекции однотипных документов (научные статьи журнала, материалы конференции) возможна автоматизация процесса извлечения метаданных. Алгоритм такой экстракции основан на анализе синтаксического уровня представления информации.

Научные статьи размечены в соответствии со стилевыми правилами, принятыми в научных журналах, и поэтому имеют относительно регулярную структуру для определенного блока электронных документов.

Математические статьи в большинстве случаев создаются с помощью систем, основанных на TeX-нотации. Но, несмотря на продвинутые возможности структурирования документа, заложенные в TeX-системы, в научных журналах, за редким исключением, используются упрощенные (с семантической точки зрения) средства структурирования. Наиболее сложными в этом плане являются архивы научных статей прошлых десятилетий, когда электронная форма документа являлась промежуточной и использовалась только для редактирования и подготовки перед печатью. Структура такого документа определяется на основе анализа шрифтового выделения и порядка следования текстовых единиц (название, автор, аннотация). Этого недостаточно для выделения ключевых слов.

Основой алгоритма структурирования журнальных статей по математике являлась обработка информации из стилевых файлов, используемых при предпечатной подготовке журнала. Название статьи, ее авторы, выходные данные, УДК определялись автоматически выделением тега, характерного для данного элемента. Создание программной среды, реализующей указанный алгоритм, позволило автоматизировать процесс структурирования электронной коллекции математического журнала.

4 Сервисы электронных математических коллекций

Как известно, сегодня поиск является самым распространенным инструментом доступа к информации в сети. По многим оценкам, поиск занимает до 50% времени работы на компьютере, а самая сложная проблема – отделение значимой информации от информационного мусора. Семантический Веб, будучи частью глобальной концепции развития интернета, имеет целью реализацию возможности машинной обработки информации и позволит рассматривать интернет в целом как глобальную базу данных. Один из акцентов этой концепции – работа с метаданными, однозначно характеризующими свойства и содержание сетевых ресурсов, вместо текстового анализа документов. Поэтому экстракция метаданных является необходимой составной частью процесса автоматизации управления электронной научной коллекцией. Вместе с тем, метаданных недостаточно для интеграции электронных коллекций в информационное пространство, в котором поиск и обработка информации программируются как машиноориентированные. В настоящее время имеется широкий набор программных средств для семантической разметки электронных документов и записи их в XML-формате, в частности, преобразования документов из TeX-нотации в MathML. Однако исходные файлы документов электронный коллекций, как правило, не удовлетворяют требованиям имеющихся пакетов и сервисов семантического преобразования из-за многообразия стилевых конструкций и отсутствия разделения на структурные элементы. Поэтому необходимым этапом становится предварительная трансформация электронных документов, обеспечивающая им структуру, общую для данной коллекции, и возможность дальнейшей автоматизированной обработки. Разработанный алгоритм трансформации электронных документов основан на синтаксическом анализе документов (см. раздел 3).

Практическая реализация описанного подхода, выполненная авторами для нескольких электронных математических коллекций, выявила дополнительные сложности, связанные с наличием авторских конструкций в электронных документах, входящих в эти коллекции. Большинство этих сложностей удается преодолеть за счет использования специализированных сервисов на всех этапах формирования электронной коллекции, в частности, электронного научного журнала (машинное взаимодействие авторов и редакции, анализ соответствия представляемых материалов заданной структуре и т. д.).

Одна из систем таких сервисов создана при автоматизации работы электронного журнала Lobachevskii Journal of Mathematics. Кратко перечислим функциональные возможности разработанной системы: вывод списка ссылок на статьи, входящие в коллекцию; вывод списка авторов статей, входящих в коллекцию; поиск по авторам, заглавиям, ключевым словам, рефератам, тексту статей. Отдельно выделим поиск по математическим формулам. Этот сервис основан на использовании технологии MathML (см., напр., [8], [9]).

Заключение

Использование технологий семантического Веба является основой интеграции электронных научных

коллекций в информационное научное пространство. Автоматизация процесса структурирования имеющихся электронных математических ресурсов создает возможность быстрого включения электронных версий математических публикаций в информационное научное пространство.

Литература

[1] Елизаров А. М., Липачев Е. К., Малахальцев М. А. Технологии Semantic Web в практике работы электронного журнала по математике // Тр. 8-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2006, Суздаль, Россия, 2006. – С. 215-218.

[2] Веселаго В. Г., Елизаров А. М., Липачёв Е. К., Малахальцев М. А. Формирование и поддержка физико-математических электронных научных изданий: переход на технологии семантического веба // В кн. «Научно-исследовательский институт математики и механики им. Н. Г. Чеботарева Казанского государственного университета. 2003 – 2007 гг.». Кол. монография под ред. А. М. Елизарова. – Казань: Изд-во Казан. ун-та, 2008. – С. 456-476.

[3] Когаловский М. Р., Паринов С. И. Семантическое структурирование контента научных электронных библиотек на основе онтологий // В сб. «Современные технологии интеграции информационных ресурсов: сборник научных трудов», 2011. – Вып. 2. – www.cemi.rssi.ru/mei/articles/kogalov11-04.pdf.

[4] Паринов С. И., Когаловский М. Р. Технология семантического структурирования контента научных электронных библиотек // Тр. 13-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2011, Воронеж, 2011. – С. 94-103.

[5] Когаловский М. Р., Хохлов Ю. Е. Стандарты XML для электронного правительства. – М.: Институт развития информационного общества, 2008. – 416 с.

[6] Когаловский М. Р., Хохлов Ю. Е. Стандарты Всемирной паутины в разработках электронного правительства. – Информационное общество: научно-аналитический журнал. – 2009. – № 2. – С. 21-32.

[7] Глухов В. А., Елизаров А. М. Проект «Научная электронная библиотека eLibrary.ru» и российские электронные журналы: новый этап развития //Тр. 8-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2006, Суздаль, Россия, 2006. – С. 203-207.

[8] Елизаров А. М., Липачев Е. К., Малахальцев М.А. Веб-технологии для математика: Основы MathML. Практическое руководство. – М.: Физматлит, 2010. – 216 с.

[9] Елизаров А.М., Липачёв Е. К., Малахальцев М.А. Языки разметки семантического веба. Практические аспекты. – http://www.ksu.ru/fpk/docs/lip_mal.pdf.

Services structuring mathematical content and integration of digital mathematical collections at scientific information space

Alexander Elizarov, Denis Zuev, Eugene Lipachev, Michael Malakhaltsev

The approach to automate the processing of scientific digital documents and convert them into structured documents is discussed. Main emphasis is placed on the features of processing of mathematical texts. Using special services which were created by the proposed method of structuring texts the large enough digital repository with periodical journal issues in mathematics and multivolume conference proceedings was performed.