Технологии управления разнородным естественнонаучным контентом на основе семантического веба♣

© А. М. Елизаров, Е. К. Липачев, М. А. Малахальцев
НИИ математики и механики им. Н. Г. Чеботарева
Казанского государственного университета
amelizarov@gmail.com

Аннотация

В работе рассматриваются подходы к формированию электронных коллекций, организации хранения и поиска данных на основе XML и других технологий семантического веба, реализованные в проекте «Научная электронная библиотека eLibrary.ru» и электронном журнале «Lobachevskii Journal of Mathematics» (http://ljm.ksu.ru). Обсуждается возможность создания электронных хранилищ нового типа, характеризующихся наличием динамических интеллектуальных связей между документами разных типов, на основе спектра языков разметки, широко применяемых в математических, химических, биологических и других предметных областях.

1 Введение

В настоящее время во всем мире, в том числе в России, идет активная работа по созданию электронных хранилищ научных документов, в частности, создаются и развиваются разнообразные электронные научные коллекции. Разработаны основные принципы организации таких коллекций и соответствующее программное обеспечение (см. обзор [1]). На этих принципах организовано большинство электронных хранилищ (например, [2 – 5]). Как правило, научные электронные коллекции представляют собой набор документов (в основном текстов статей и книг) и их «библиографических» описаний, построенных на основе языка XML. Поэтому научные электронные коллекции уже сейчас позволяют организовать поиск не только по текстовой информации, но и по XML-описанию.

Одним из первых в России в области управления электронной научной информацией был проект «Научная электронная библиотека eLibrary.ru» (см., например, [5]). В рамках этого проекта в части разработки методов управления электронным научным контентом нами были предложены технологические решения, базирующиеся на идеях семантического веба [2], [6]. Эти подходы были реализованы в системе управления математическим контентом электронного журнала «Lobachevskii Journal of Mathematics» (http://ljm.ksu.ru) [2].

2 Управление научным контентом в НЭБ

Для структурирования макетов печатных изданий в электронной коллекции, созданной в рамках проекта «Научная электронная библиотека elibrary.ru» (НЭБ), была разработана программная среда, в основу которой положен алгоритм выделения элементов текста и присвоения им меток полей собственного XML-формата, названного Sarcticle (подробности см. в [5]). 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, выпуске и статье (основная информация файла). Большинство полей может дублироваться на нескольких языках с целью более удобного представления для разных пользователей конечной информации в электронной библиотеке.

Задача подготовки библиографических материалов, включаемых в различные индексы научного цитирования, решается с помощью программного модуля, производящего автоматическое структурирование (разбор по полям формата) списков литературы и сносок. В этом модуле учтены требования ГОСТ 7.1-84 «Библиографическое описание документа».

Названное программное обеспечение позволяет работать с большинством форматов (в том числе, html, PageMaker, .pdf, .doc) и максимально автоматизирует процесс структуризации текста макетов. Одной из основных задач, которая была решена при разработке программы разметки текстов электронных журналов, состояла в организации обработки и отображения математических и химических формул, диакритических, математических и других специа-

Труды 11ⁱᵉ Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.

лизированных знаков и символов.

Использование формата Sarcticle позволило создать программный комплекс, поддерживающий расширенный поиск статей по журнальной базе, в частности, возможен поиск по следующим параметрам: авторы, названия статей, аннотации (рефераты), ключевые слова, слова из полных текстов статей, библиографические описания источников.

Программное обеспечение разметки электронных журналов прошло успешное тестирование в ряде редакций научных журналов. По разработанной технологии в Научной электронной библиотеке eLibrary.ru было размещено более 200 научных журналов по разным специальностям, в том числе «Успехи физических наук», «Успехи химии», «Биология моря», «Ученые записки Казанского университета», «Известия вузов. Авиационная техника», «Известия вузов. Математика», «Известия вузов. Радиофизика», «Казанский медицинский журнал», «Тихоокеанская геология» и ряд других. Большинство из названных журналов до этого момента времени не было представлено в интернете.

Указанные технологии и возможности НЭБ характерны для современных электронных библиотек, однако развитие информационных технологий делает возможным переход в организации электронных коллекций на качественно новый уровень на основе технологий семантического веба.

3 Технологии семантического веба и электронные хранилища нового типа

Согласно программным документам консорциума W3C [7], семантический веб – это «... расширение традиционного веба в направлении существенно лучшего определения смысла информации, позволяющего компьютерам и людям эффективнее выполнять совместную работу. Мы хотим, чтобы данные в вебе были определены и связаны ссылками так, чтобы их можно было легче находить, интегрировать, автоматизировать и повторно использовать в различных приложениях, ... чтобы данные были разделяемыми и могли обрабатываться как автоматизированными средствами, так и людьми». Конечная цель этого проекта состоит в создании такой среды, в которой программные агенты смогут динамически обнаруживать и опрашивать ресурсы, а затем взаимодействовать с ними. Такие агенты должны уметь справляться с возникающими виртуальными проблемами интеллектуализированной среды, обнаруживать новые факты и выполнять изощренные задания, получаемые от людей (см., например, [8]).

Основные цели семантического веба, приведенные выше, естественным образом переносятся на электронные коллекции и электронные библиотеки, позволяя говорить об *электронных хранилищах нового типа.* С нашей точки зрения, можно выделить следующие характеристики, присущие таким электронным хранилищам.

1) *Разнородность контента* – понимается как разнообразие предметных областей (математика, физика, химия, биология, геология и т. д.), разнообразие как типов документов (научные статьи, результаты наблюдений и экспериментов, программные продукты), так и форматов (текстовый, графический, звуковой, видео). Кроме того, структура документов хранилища может быть отличной от традиционных научных статей. Например, документ по астрономии может включать одновременно текст, математические расчеты, данные наблюдений и программы обработки этих данных.

2) *Возможности интеллектуального поиска по специализированным документам*. Сегодня для подавляющего большинства коллекций электронных документов основная задача состоит в предоставлении пользователю информации по запросу, который формирует сам пользователь, например, указав предметную область и ключевые слова. В электронном хранилище нового типа должны быть установлены *динамические интеллектуальные связи* между документами различных типов, позволяющие предоставлять пользователю информацию из разных областей и разных типов. Простой пример: при запросе о дифференциальном уравнении определенного типа должна выдаваться информация не только о математических результатах, связанных с этим уравнением, но и сведения о его применениях, вычислительных экспериментах, связанных с ним, и т. д. Подчеркнем, что в документах, где данное уравнение используется, оно может быть не упомянуто как ключевой термин и даже не названо, однако система электронного хранилища сама должна установить необходимые связи. Поэтому в целом система управления электронным хранилищем должна предоставлять пользователю *структурированный комплекс документов*.

3) *Новый тип интерфейса*. Интерфейсы имеющихся хранилищ ориентированы на взаимодействие с человеком. Хранилище нового типа, согласно принципам семантического веба, должно иметь интерфейс, приспособленный для взаимодействия на программном уровне («машинно-ориентированный подход» [9]).

Электронные хранилища нового типа соответствуют задачам современной науки, где на первый план вышли междисциплинарные исследования. Создание таких хранилищ возможно благодаря тому, что к настоящему времени для большинства естественнонаучных предметных областей уже созданы специализированные языки разметки.

Язык разметки химических формул CML (Chemical Markup Language) разработан как часть проекта Open Molecule Foundation. С помощью CML записывается информация о молекулярных структурах, химических реакциях, спектрах, неорганических кристаллах, объектах квантовой химии. Для создания и обработки CML-файлов можно использовать уже созданные программные средства, предназначенные для работы с информацией в формате XML [10 – 12].

Для представления математических формул в

рамках семантического веба используется технология MathML [13]. Разработка этого языка ведется консорциумом W3C с 1998 года. Фактически он уже стал стандартом представления математической информации в электронной форме [14, 15].

Для хранения и электронного обмена математическими моделями применяется язык CellML (Cell Markup Language). В частности, он широко используется в биологическом моделировании. Отметим, что CellML поддерживает спецификацию MathML.

Для описания свойств материалов можно использовать язык Materials Markup Language (MatML). Подробное его описание можно найти на сайте www.matml.org.

Географическим сообществом используется язык Geography Markup Language (GML) (информацию о нем можно найти на сайтах http://www.opengis.net/gml/ и http://www. opengeospatial.org/standards/gml, http://schemas.opengis.net/gml/).

Отметим, что созданы языки разметки и для других предметных областей, причем названия языков, как правило, отражают их назначение: Business Rules Markup Language (BRML), Geography Markup Language (GML), Finite Element Modeling Markup Language (femML), Ink Markup Language (InkML), Mathematics Education Markup Language (MeML), Materials Markup Language (MatML), Numerical Data Markup Language (NDML), Relational-Functional Markup Language (RFML), Robotic Markup Language (RoboML), Voice Extensible Markup Language (VoiceXML). В работе [16] предложена классификация уже созданных языков разметки в виде карты языков XML.

Важно, что технологии семантического веба обеспечивают стандартную процедуру создания языка разметки, адаптированного к определенной предметной области [12], что позволяет гибко настраивать структуру хранилища для включения информации из новых предметных областей.

Структура естественнонаучного электронного хранилища нового типа основана на имеющихся в настоящий момент времени языках разметки, широко применяемых в математических, химических, биологических и других предметных областях. Общая природа этих языков позволяет использовать уже отлаженные технологии семантического веба для управления разнородным контентом естественнонаучного электронного хранилища.

Заключение

Таким образом, имеющиеся на сегодняшний день технологии семантического веба, в частности, разработанные языки разметки естественнонаучной информации позволяют решить задачу управления разнородным контентом в электронных хранилищах.

Литература

[1] Когаловский М. Р. Тенденции развития технологий управления информационными ресурсами в электронных библиотеках// Тр. 8-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2006, Суздаль, Россия, 2006.

[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] Бархатов А. В., Вдовицын В. Т., Луговая Н. Б., Сорокин А. Д. Электронные научные информационные ресурсы для поддержки инвестиционной деятельности в регионе. – www.kareliainvest.ru/file.php/id/f3770/name/STATIA_Inf_res.doc.

[4] Голосов Ю. И., Брагина Г. А., Пржиялковская М. Н. Электронные документы научно-технической информации в системе ВНТИЦ// Тр. 10-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

[5] 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.

[6] 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.

[7] W3C Semantic Web Activity Statement. – http://www.w3.org/2001/sw/Activity.

[8] Hendler J. Agents and the semantic web// IEEE Intelligent Systems J. – 2001. – V. 16, No 2. – P. 30-37.

[9] Berners-Lee T. Semantic web road map. – http://www.w3.org/DesignIssues/Semantic.html; Рус. перевод: http://gridclub.ru/library/publication. 2007-04-23.2195467714/view.

[10] Jirat J. Chemical Markup Language 1.0 reference with examples. – http://www.zvon.org/xxl/CML1.0

[11] Amies A. Tools for working with Chemical Markup Language. – http://www.medicalcomputing. net/cmltools.html.

[12] Елизаров А. М., Липачёв Е. К., Малахаль-

цев М. А. Языки разметки семантического веба. Практические аспекты. – http://www.ksu.ru/fpk/docs/lip_mal.pdf.
[13] Mathematical Markup Language (MathML) Version 2.0 (Second Edition). – http://www.w3.org/TR/2003/REC-MathML2-20031021/.
[14] Елизаров А. М., Липачёв Е. К., Малахальцев М. А. Основы MathML Представление математических текстов в Internet. Практическое руководство. – Казань: Изд-во Казан. матем. общества, 2004. – 60 с. – www.ksu.ru/.
[15] Елизаров А. М., Липачёв Е. К., Малахальцев М. А. Основы MathML Представление математических текстов в Internet. – Казань, 2008. – 101 с. – http://www.niimm.ksu.ru/data/preprints/.
[16] Лозовюк А. Комета по имени XML. – http://www.marketer.ru/.

Management technology for multi-discipline scientific content based on Semantic Web

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

In the present paper we consider approaches to formation of electronic collections, organization of data storage and search on the base of XML and other Semantic Web technologies, which were used in the project «Scientific electronic library eLibrary.ru» and in the electronic journal «Lobachevskii Journal of Mathematics» (http://ljm.ksu.ru). We propose to design new type electronic storages which are characterized by dynamical knowledge-based relations between documents of different types with the use of the variety of markup languages widely applied in mathematics, chemistry, biology and other areas of science.

♣ Работа поддержана РГНФ (проект № 07–01–12146) и РФФИ (проект № 09–07– 12059-офи_м)