Разработка и развитие технологии публикации и поиска документов в электронных коллекциях

Вдовицын В.Т., Лебедев В.А., Луговая Н.Б., Сорокин А.Д., Старкова В.Г.

Институт прикладных математических исследований КарНЦ РАН, г. Петрозаводск
e-mail: vdov@krc.karelia.ru

Аннотация

В статье представлены результаты, связанные с разработкой, развитием и реализацией технологии формирования, сопровождения и поиска электронных коллекций научных информационных ресурсов с использованием возможностей платформы XML, которые были получены в рамках создания электронной библиотеки научных информационных ресурсов КарНЦ РАН (http://dl.krc.karelia.ru). При этом особое внимание уделено вопросам разработки и применения онтологии для построения более эффективных алгоритмов поиска данных в коллекциях, а также вопросам развития программных сервисов и информационного содержания электронной библиотеки с учетом опыта эксплуатации системы и появления новых задач.

1. Введение

В настоящее время остается актуальной проблема формирования и эффективного использования электронных научных информационных ресурсов для поддержки процессов проведения фундаментальных исследований, образования и инновационной деятельности. Одним из подходов к ее решению являются исследования и разработки в области создания электронных коллекций и электронных библиотек [1,2].

В данной работе представлены результаты, связанные с разработкой, развитием и реализацией в рамках создания научной электронной библиотеки Карельского научного центра Российской академии наук (ЭБ КарНЦ РАН) – http://dl.krc.karelia.ru информационной технологии формирования, сопровождения и использования электронных коллекций научных информационных ресурсов с применением возможностей платформы XML [7,8].

Труды 8ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2006, Суздаль, Россия, 2006.

В отличие от большинства существующих электронных библиотек, содержащих, как правило, электронные версии полнотекстовых печатных изданий, в нашем случае основной акцент сделан на разработку и реализацию технологии публикации в сети Интернет электронных коллекций научных информационных ресурсов, создаваемых на основе результатов многолетних исследований ученых. При этом следует отметить, что научные коллекции обладают рядом особенностей, в частности, они "более динамичны по структуре и составу информационных ресурсов по сравнению, например, с музейными коллекциями, а информационные потребности исследователей более разнообразны чем, например, относительно стабильные информационные потребности управленческих работников" [9].

Информационная технология формирования, сопровождения и поиска научных электронных коллекций разрабатывается нами с учетом следующих основных положений. Во-первых, специалисты-предметники должны, на наш взгляд, в максимальной степени самостоятельно осуществлять формирование и сопровождение своих коллекций в информационной системе с учетом общепринятых в их научной среде стандартов и в рамках определенного для них в ЭБ КарНЦ РАН регламента работы. Это должно способствовать улучшению качества и достоверности информационного содержания электронных документов коллекции и повышению заинтересованности специалистов в конечном результате своей работы. Во-вторых, сервисы ЭБ должны включать удобные для специалистов-предметников программные средства поддержки процессов формирования, публикации и сопровождения своих коллекций с учетом разграничения их полномочий и защиты информационных ресурсов от несанкционированного доступа. В-третьих, пользователи ЭБ должны иметь эффективные средства для доступа к нужной информации по запросам.

Разработка, развитие и реализация предлагаемой информационной технологии в рамках создания ЭБ КарНЦ РАН будет способствовать в первую очередь достижению следующих основных целей: созданию

серии тематических коллекций в рамках инвентаризации природных, социально – экономических и культурно – исторических ресурсов Карелии; публикации результатов исследований по их эффективному использованию и охране, рекомендаций по сохранению и приумножению этих ресурсов, а также организации новых форм проведения совместных научных исследований с использованием возможностей сети Интернет.

Данная работа является продолжением серии публикаций, связанных с разработкой, развитием и реализацией технологии формирования и использования электронных коллекций научных информационных ресурсов в среде XML. В разделе 2 приведено краткое описание основных решений по разработке и реализации текущей версии ЭБ КарНЦ РАН. В разделе 3 описан подход к разработке нового алгоритма поиска данных в электронной библиотеке на основе применения онтологии, а в разделе 4 представлены направления развития программных сервисов и информационного содержания электронной библиотеки с учетом опыта ее эксплуатации и появления новых задач.

2. Описание технологии и основных программных сервисов

В рамках разрабатываемой технологии процесс формирования и сопровождения каждой предметной коллекции научных информационных ресурсов проходит по следующей схеме. Для создания новой коллекции формируется группа специалистов, включающая предметного администратора коллекции, авторов документов и экспертов. Эти категории пользователей регистрируются в системе и получают определенный “объем” прав работы с данной коллекцией. Администратор и авторы документов предметной коллекции разрабатывают паспорт описания исследуемых объектов, в котором структурируется разнородная (текст, графика, аудио и т. п.) научная информация об этих объектах. На основе разработанного паспорта технический администратор системы формирует DTD-определение структуры описания класса соответствующих XML - документов новой коллекции, а также “прописывает” новую коллекцию в системе. Предметный администратор коллекции совместно с авторами документов и с учетом мнений экспертов коллекции описывает общие свойства предметной коллекции с использованием атрибутов стандарта Дублинского ядра (DC, Dublin Core). После этого происходит ввод (корректировка) документов в коллекцию и обсуждение их информационного содержания с экспертами на форуме ЭБ.

Основные программные сервисы ЭБ КарНЦ РАН предназначены для поддержки процессов публикации и сопровождения документов коллекций, а также для поиска данных в коллекциях по запросам пользователя.

Для поддержки процессов публикации и сопровождения документов коллекций разработана специальная технология, обеспечивающая дружественный интерфейс при заполнении полей документов самими специалистами-предметниками, а также автоматическую верификацию целостности документов и формирование на основе заданного DTD – определения коллекции соответствующих действительных XML-документов.

Поиск данных в ЭБ КарНЦ РАН происходит в два этапа. Сначала осуществляется поиск нужной коллекции среди других коллекций электронной библиотеки, а затем - поиск документа в выбранной коллекции по запросу пользователя. При этом поиск коллекции осуществляется пользователем как при помощи специально разработанного рубрикатора, который формируется на основе ГРНТИ и в соответствии с направлениями проводимых в центре исследований, так и по базе метаданных, описывающей общие свойства коллекций в стандарте DC. Запросы на поиск документов в выбранной коллекции формируются пользователем с помощью специально разработанных интерфейсных форм путем указания значений полей искомого документа. При этом список полей, по которым возможен отбор документов, специфичен для каждой коллекции.

Для реализации разработанных программных сервисов используются технологии XML, DTD, XSLT, объектная модель документа – DOM, а также традиционные информационные технологии - СУБД MySQL и язык PHP.

Более подробное описание вопросов реализации программных сервисов ЭБ КарНЦ РАН представлено в работах [7,8].

В настоящее время ведутся работы как по созданию новых и совершенствованию ранее разработанных программных сервисов, так и по развитию информационного содержания (контента) электронной библиотеки. Одним из таких направлений развития системы является разработка нового сервиса тематического поиска данных в коллекциях библиотеки с использованием онтологии.

3. Применение онтологии для организации тематического поиска данных

Развитие контента ЭБ КарНЦ РАН предполагает создание набора коллекций, представляющих сведения о состоянии и перспективах использования природных ресурсов, исторического и культурного наследия Карелии. Учитывая тот факт, что

информационные потребности исследователей более разнообразны, чем, например, информационные потребности музейных работников или управленцев, представляется актуальным разработка нового программного сервиса поиска документов из одной или нескольких разных коллекций на заданные темы, т.е. организация тематического поиска документов в коллекциях электронной библиотеки. При этом предполагается, что поисковый запрос может содержать множество терминов, находящихся в конъюктивно - дизъюнктивных связях, а тематический подбор документов может потребовать исполнения серии запросов к разным коллекциям.

В настоящее время реализованные в ЭБ КарНЦ РАН поисковые сервисы предоставляют ограниченные возможности поиска документов в коллекциях, в основном по названиям систематических группировок (классы, отряды, семейства, виды) организмов, минералов и т.п. При этом названия терминов набираются пользователем вручную, что часто бывает затруднительным, особенно в тех случаях, когда имеет место синонимия. Между тем в имеющихся наборах документов возможен тематический поиск по различным аспектам. Например, интересным может быть подбор видов лекарственных растений для лечения некоторой болезни, определение возможного флористического состава местообитаний и др.

Для формирования тематических поисковых запросов на наш взгляд целесообразно использовать онтологии предметных областей. При этом онтология определяется как спецификация концептуального представления предметной области, соответствующая поставленным задачам (T. Gruber, M. Ushold) [12].

В соответствии с принятой схемой формирования электронной коллекции в процессе разработки структуры документа коллекции одновременно определяются словари терминов, отображающих классификации, строение и связи объектов рассматриваемой предметной области. Для решения этой последней задачи применяется методология, сходная с предложенной нами методологией формирования онтологий [10]. Таким образом, для создания конкретной онтологии необходимо лишь зафиксировать результат проделанной работы с помощью разработанной нами технологии построения онтологий, а также технологии построения на основе онтологий поисковых запросов и алгоритма поиска релевантных документов через базу данных индексов документов [11].

Как известно, поиск релевантных запросу документов по базе данных индексов эффективнее по времени полнотекстового поиска. Однако в этом случае возникает задача индексирования документов, которая при наличии онтологии может быть выполнена автоматически при помощи специально разработанного программного сервиса.

Ввиду многообразия предметных областей, планируемых к представлению в ЭБ, и возможных различиях назначения коллекций, затруднительно построить онтологии предполагаемых в ЭБ предметных областей с учетом актуальности поисковых задач. В связи с этим с учетом предполагаемых информационных потребностей пользователей представляется целесообразным разрабатывать конкретные онтологии для предметных коллекций и соответствующие алгоритмы индексирования.

Алгоритм индексирования построен на использовании соответствия структуры документов и структуры онтологий указанных предметных областей. Состав терминов онтологий предусматривает возможность тематического подбора документов по систематике организмов, экологическим характеристикам и возможному их использованию.

При построении данного сервиса основная трудность заключалась в реализации установления соответствия многословных терминов в текстах документов и в онтологии ввиду падежной изменчивости морфологии слов в русском языке.

В настоящее время разработаны онтологии и алгоритмы индексирования для коллекций: «Виртуальная флора Карелии», «Афиллофороидные грибы Карелии», «Млекопитающие Карелии». Целесообразность поиска документов с использованием онтологий определяется следующими факторами. Во-первых, пользователю не приходится набирать вручную многочисленные термины и при этом не нужно задумываться о возможных синонимах. Во-вторых, благодаря наличию онтологии пользователь получает возможность получить сжатое представление о содержании коллекции и на этой основе осознанно строить запросы на подбор документов по интересующим его темам. Например, он может подбирать виды растений по местообитаниям определенного типа, или лекарственных растений для лечения определенных болезней, или декоративных, или кормовых растений, однолетних и многолетних и т.д. О млекопитающих он может узнать об отношениях в зооценозах, какие враги, паразиты или болезни свойственны данному виду и т.д.

Таким образом, применение онтологий для тематического поиска документов обогащает возможности работы пользователей с информационным содержанием ЭБ.

4. Развитие программных сервисов и информационного содержания

Разработка и развитие программных сервисов ЭБ КарНЦ РАН в нашем случае представляет собой “итеративный” процесс, который осуществляется как с учетом приобретения опыта эксплуатации системы, так и с учетом появления новых задач.

Одним из таких примеров является создание предметной электронной коллекции документов «Местообитания Восточной Фенноскандии» - http://dl.krc.karelia.ru/search_doc.html?url=biotopy, работа над которой привела к необходимости разработки нового программного сервиса автоматического формирования XML – документов коллекции из локальных баз данных. Комплексные исследования биотопов Карелии проводились учеными центра в течение ряда многих лет. При этом описания биотопов, полученные в результате полевых экспедиций, заносились в специально разработанную в среде СУБД MS Access информационную систему. При создании коллекции документов по разработанной нами технологии возникла задача автоматического формирования на основе информации из локальных баз данных в среде СУБД MS Access корректных XML – документов, соответствующих заданному DTD – определению их структуры. Для решения задачи преобразования базы данных «Местообитания Восточной Фенноскандии» из формата Microsoft Access в XML-формат электронной библиотеки КарНЦ РАН была разработана на языке Java программная система DbReader (разработчик – Крышень М. М.), которая обрабатывает заданные шаблоны текстовых файлов и заменяет найденные в них инструкции на результаты их выполнения. При этом заданные инструкции позволяют формировать SQL-запросы, выполнять их и обрабатывать полученные данные. Разработанный программный сервис позволяет (после соответствующей настройки) автоматически формировать корректные (действительные) XML – документы в коллекциях ЭБ КарНЦ РАН и является универсальным средством для представления информации из баз данных в текстовых форматах. Этот сервис позволяет существенно уменьшить трудозатраты специалистов при создании электронных коллекций XML – документов на основе информации из локальных баз данных.

Другим примером применения создаваемой нами информационной технологии формирования электронных коллекций научных информационных ресурсов является разработка по заказу Министерства экономического развития Республики Карелия предметной коллекции «Минерально – сырьевые ресурсы Республики Карелия» - http://dl.krc.karelia.ru/search_doc.html?url=minerals [3]. Коллекция предназначена в первую очередь для информационной поддержки инвестиционной деятельности в регионе за счет формирования, сопровождения и эффективного использования электронных информационных ресурсов, создаваемых на основе результатов многолетних научных исследований специалистов из Карельского научного центра РАН, Петрозаводского государственного университета и других научных и проектных организаций. Структура и информационное содержание документов коллекции разработаны с учетом рекомендаций по составлению бизнес-планов, а также информационных потребностей потенциальных инвесторов, и включают следующие основные разделы: название и фото минерала; общие сведения о минерале (включая инновационные аспекты использования); анализ рынков; описание минерально – сырьевой базы Карелии (включая карты – схемы размещения месторождений); описание общей оценки готовности месторождений Карелии к инвестированию, а также ссылки на подобные источники информации и контактную информацию.

Наиболее важными с нашей точки зрения вопросами при создании такого рода предметных электронных коллекций является обоснованный выбор структуры и информационного содержания документов коллекции, а также программная поддержка совместной работы специалистов по созданию и сопровождению электронной коллекции. Это связано в первую очередь с тем, что работа по формированию такой коллекции требует привлечения различных специалистов, которые должны оперативно и согласованно корректировать и пополнять документы коллекции новыми данными. В этом плане разработанные в рамках создания ЭБ КарНЦ РАН программные сервисы для поддержки процессов формирования и сопровождения коллекций являются на наш взгляд удобным средством для проведения такой работы.

В процессе работы над данной коллекцией выявилась необходимость как в получении обычных статистических данных о посетителях коллекции, так и в разработке инструмента для изучения “поведения” пользователей при работе с документами коллекции. В частности, специалистов органов государственного управления может заинтересовать информация об обнаружении у определенных категорий пользователей коллекции устойчивой зависимости в “поведении”, характеризующей проявление их “интереса” к определенным группам полезных ископаемых.

Таким образом, при создании данной коллекции осуществляется разработка новых программных сервисов системы по следующим основным направлениям. Во-первых, для получения статистики посещения коллекции разработан программный сервис регистрации пользователей коллекции. Во-вторых, разрабатывается структура базы данных, которая включает как

регистрационные данные о посетителях коллекции, так и данные о проведенных ими логических сессиях. При этом предполагается разработка специального приложения, которое на основе анализа соответствующих Log-файлов будет автоматически пополнять базу данных новыми данными [5]. В-третьих, в перспективе планируется применить к этой базе данных разработанные нами на основе алгоритмов поиска логических зависимостей и ассоциативных правил программные средства системы DMiner [4,6]. Это позволит использовать возможности технологии KDD (Knowledge Discovery in Databases) для проведения более “глубокого” анализа посещений коллекций ЭБ пользователями с целью выявления в их “поведении” определенных закономерностей.

5. Заключение

В статье представлены результаты, связанные с разработкой, развитием и реализацией технологии формирования, сопровождения и поиска электронных коллекций научных информационных ресурсов с использованием возможностей платформы XML, которые были получены в рамках создания электронной библиотеки КарНЦ РАН - http://dl.krc.karelia.ru.

Главная цель разработки такой технологии заключается в том, чтобы обеспечить исследователям возможность оперативной публикации в сети Интернет своих научных результатов в виде электронных коллекций документов, а также организовать их сопровождение и эффективный поиск данных по запросам пользователей.

В связи с тем, что информационные потребности исследователей разнообразны и динамичны, процесс создания такой системы в нашем случае носит “итеративный” характер. Это означает, что состав сервисов и их функции определяются и уточняются в процессе изучения опыта эксплуатации системы, а также появления новых информационных потребностей у пользователей электронной библиотеки.

Одним из таких примеров является разработка нового программного сервиса тематического поиска данных в коллекциях ЭБ, основанного на применении онтологии. Реализация этого сервиса позволит пользователям электронной библиотеки строить запросы на подбор документов из разных коллекций по интересующим его темам.

В статье также представлены новые коллекции научных информационных ресурсов, разработка которых привела к необходимости построения программного сервиса для автоматического формирования действительных XML – документов коллекций из локальных баз данных. Это позволило сократить трудозатраты исследователей на осуществление “рутинной” информационной работы по формированию документов коллекции «Местообитания Восточной Фенноскандии» на основе ранее созданных баз данных.

При создании предметной коллекции «Минерально – сырьевые ресурсы Республики Карелия» возникла задача анализа поведения пользователей при их доступе к документам коллекции. Актуальность этой задачи определяется в первую очередь потребностями специалистов органов государственного управления, например, в плане отслеживания “интереса” потенциальных инвесторов к тем или иным месторождениям. В данной работе описан перспективный подход к построению программного сервиса для проведения анализа посещений коллекций пользователями ЭБ с применением технологии KDD с целью обнаружения в их “поведении” определенных закономерностей.

Данная работа выполняется при поддержке РФФИ (грант № 05-07-90077).

Литература

[1] http://www.elbib.ru/ - информационный портал "Российские электронные библиотеки".

[2] Акимов С. И., Елизаров А. М., Ершова Т. В., Когаловский М. Р., Федоров А. О., Хохлов Ю. Е. Научно-методическая поддержка разработки научных электронных библиотек. //Российский научный электронный журнал "Электронные библиотеки". 2005.-Том 8-Выпуск 1.

[3] Бархатов А. В., Вдовицын В. Т., Сорокин А. Д., Луговая Н. Б. Электронные научные информационные ресурсы для поддержки инвестиционной деятельности в регионе. //Информационные ресурсы России. 2006. (принята к печати).

[4] А. А. Бедорев, Ю. В. Чуйко Применение алгоритмов поиска регулярных эпизодов для анализа посещений Web – сайта. //Труды Института прикладных математических исследований. Методы математического моделирования и информационные технологии. – Петрозаводск: Карельский научный центр РАН, 2002. С. 153-169.

[5] Николай Бузикашвилли Поисковое поведение пользователя Яндекса (анализ веблогов). //Сб. статей: “Интернет-математика 2005. Автоматическая обработка веб-данных”. Москва, 2005. С. 95-120.

[6] В.Т. Вдовицын, Г.М. Керт, Н.Б. Луговая, Ю.В. Чуйко Применение алгоритмов поиска логических зависимостей для решения задач в области топонимики. Обозрение прикладной и промышленной математики. Том 10. Вып.2, 2003 г., с.387-388.

[7] В.Т. Вдовицын, А.Д. Сорокин, Луговая Н.Б. Электронная библиотека научных информационных ресурсов КарНЦ РАН: состояние и перспективы развития //Труды Шестой Всероссийской научной конференции

"Электронные библиотеки: перспективные методы и технологии, электронные коллекции ", Пущино, 29 сентября-1 октября 2004 г. С.41-46 .
[8] Вдовицын В. Т., Сорокин А. Д., Луговая Н. Б. Развитие программных сервисов и контента ЭБ КарНЦ РАН. //Труды Седьмой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" (Ярославль, 4-6 октября 2005 года). - Ярославль: Ярославский государственный университет им. П. Г. Демидова, 2005. с. 92-97.
[9] Когаловский М. Р. Систематика коллекций информационных ресурсов в электронных библиотеках. //Программирование. № 3. 2000 г., с 31-52.
[10] Лебедев В. А., Старкова В. Г., Брагин С. В. Представление онтологии научной коллекции “Водные ресурсы региона”. // Труды Шестой Всероссийской научной конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции “, Пущино, 29 сентября-1 октября 2004 г. С. 86-92.
[11] Лебедев В. А., Старкова В. Г., Брагин С. В. Применение онтологии для ведения и доступа к данным коллекций “Природные ресурсы региона”. //Труды Седьмой Всероссийской научной конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции” (Ярославль, 4-6 октября 2005 года). – Ярославль: Ярославский государственный университет им. П. Г. Демидова, 2005. с. 87-91.
[12] Россеева О. И., Загорулько Ю. А. Организация эффективного поиска на основе онтологий. http://www.dialog-21.ru/archive_article.asp?param=7029&y=2001&vol=6078/2001.

Implementing and developing of the technology for publising and searching of documents in digital collections

Vladimir Vdovitsyn, Victor Lebedev, Natalya Lugovaya, Anatoly Sorokin, Valentina Starkova

The paper presents results on the development, advancement and implementation of a technology for formation, maintenance and searching of digital collections of scientific information resources using the capacities of the XML platform obtained while creating the digital library of scientific information resources of Karelian Research Centre, Russian Academy of Science (http://dl.krc.karelia.ru). The focus is on creation and application of the ontology for constructing most efficient data search algorithms, as well as on problems of further developing software services and information contents of the digital library with regard to the experience gained from using the system and emergence of new tasks.