*Развитие программных сервисов и контента ЭБ КарНЦ РАН**

© Вдовицын В.Т., Сорокин А.Д., Луговая Н.Б.

Институт прикладных математических исследований КарНЦ РАН
vdov@krc.karelia.ru

Аннотация

В статье представлены текущие результаты работы по созданию и развитию программных сервисов и информационного содержания (контента) электронной библиотеки научных информационных ресурсов Карельского научного центра Российской академии наук (ЭБ КарНЦ РАН) – http://dl.krc.karelia.ru.

1 Введение

Проблема формирования и эффективного использования электронных научных информационных ресурсов для поддержки процессов проведения фундаментальных исследований, образования и инновационной деятельности в настоящее время остается актуальной [1,2,3]. Одним из перспективных направлений исследований и разработок в этой области является создание электронных (цифровых) коллекций научных информационных ресурсов и научных электронных (цифровых) библиотек (ЭБ, Digital Library) [4,5,6].

Карельский научный центр Российской академии наук (КарНЦ РАН) представляет собой многопрофильное научное учреждение, включающее в свой состав 7 институтов (биологии, леса, геологии, водных проблем севера, экономики, прикладных математических исследований, языка, литературы и истории), а также ряд вспомогательных подразделений при Президиуме центра. За 60-летнюю историю своего существования ученые центра провели ряд исследований и накопили значительное количество научных информационных ресурсов, связанных в первую очередь с изучением Европейского севера России. Систематизация и структуризация этой информации и представление ее в Интернет с учетом интеграции в Единое научное информационное пространство Российской академии наук является для нас актуальной проблемой [7]. В качестве перспективного подхода к решению этой проблемы мы рассматриваем возможность создания ЭБ научных информационных ресурсов в КарНЦ РАН, которая потенциально может стать универсальным хранилищем всей электронной научной информации, создаваемой в центре. При этом главными функциями такой библиотеки является публикация всех видов научной информации и обеспечение интеграции разнородных электронных научных информационных ресурсов, а также обеспечение ее сохранности и доступности через Интернет [8].

В отличие от большинства электронных библиотек, содержащих, в основном, электронные версии полнотекстовых изданий (таких, например, как широко известная научная электронная библиотека – http://www.elibrary.ru и др.), при построении ЭБ научных информационных ресурсов КарНЦ РАН мы преследуем следующие основные цели [6]:

- обеспечить научным сотрудникам центра возможность публикации в Интернете своих результатов исследований не только в виде электронных версий научных публикаций, но и в виде тематических коллекций документов;
- обеспечить оперативный доступ к необходимым электронным информационным ресурсам, соответствующим тематике исследований центра;
- предотвратить утрату ценных научных коллекций для последующих поколений ученых;
- способствовать научному сотрудничеству коллективов ученых путем создания новых технологий проведения научных исследований (например, путем организации “виртуальных лабораторий”).

Процесс создания такой информационной системы должен осуществляться, на наш взгляд, итеративно, по мере накопления как собственного опыта, так и опыта привлекаемых к этой работе специалистов – предметников, и базироваться на современных достижениях в области построения научных электронных библиотек. При этом должны быть системно проработаны и решены различные аспекты создаваемой системы, которые

Труды 7ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL’2005, Ярославль, Россия, 2005.

варьируются от исследования и разработки технологических проблем построения ЭБ, до вопросов эффективной реализации соответствующих программных сервисов и создания конкретных предметных коллекций.

В данной работе рассматриваются вопросы, связанные с разработкой, развитием и реализацией программных сервисов и информационного содержания (контента) электронной библиотеки научных информационных ресурсов КарНЦ РАН - http://dl.krc.karelia.ru [9,10]. В разделе 2 приводится краткое описание основных решений по разработке и реализации текущей версии системы, в разделе 3 представлена разработанная процедура поиска и сортировки XML – документов в коллекциях электронной библиотеки, а в разделе 4 – описаны основные направления развития информационного содержания ЭБ КарНЦ РАН.

2 ЭБ научных информационных ресурсов КарНЦ РАН

Электронная библиотека научных информационных ресурсов Карельского научного центра Российской академии наук (ЭБ КарНЦ РАН) предназначена для формирования, хранения и многоцелевого использования коллекций электронных научных информационных ресурсов, создаваемых учеными центра на основе результатов многолетних исследований, и предоставления к ним доступа через сети Интернет. К числу таких коллекций в первую очередь мы относим: предметные коллекции по различным областям науки, исследования по которым ведутся в КарНЦ РАН; списки ссылок на другие научные Интернет – ресурсы, соответствующие тематике проводимых в центре исследований, а также коллекции электронных версий научных публикаций ученых центра.

Разработанная нами технология формирования и сопровождения предметных коллекций в ЭБ КарНЦ РАН учитывает следующие основные особенности. Во-первых, специалисты-предметники должны, на наш взгляд, в максимальной степени самостоятельно осуществлять формирование, публикацию и сопровождение своих коллекций с учетом общепринятых в их научной среде стандартов и в рамках определенного для них регламента работы. Это способствует улучшению качества и достоверности информационного содержания документов и повышает заинтересованность специалистов в конечном результате работы. Во-вторых, сервисы ЭБ должны включать удобные для специалистов-предметников программные средства автоматизации 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 удобные и эффективные средства для доступа к нужной информации по запросам.

Процесс формирования, публикации и сопровождения предметной коллекции научных информационных ресурсов в нашем случае происходит по следующей схеме. Для создания новой коллекции формируется группа специалистов, включающая 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 единый паспорт описания объектов, в котором структурируется разнородная (текст, графика, аудио и т. п.) научная информация об описываемых объектах. Эти паспорта обсуждаются и согласовываются с экспертами предметной коллекции. На основе разработанных паспортов формируется DTD-определение структуры описания класса соответствующих XML - документов новой коллекции. Администратор коллекции совместно с авторами документов и с учетом мнений экспертов коллекции описывает общие свойства предметной коллекции на основе 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 их реализации использованы технологии XML, DTD, XSLT, язык Java и Java API for XML, объектная модель документа – DOM, а также традиционные информационные технологии - СУБД MySQL и язык PHP. Выбор XML-платформы (http://www.w3.org/) в качестве основы для построения ЭБ КарНЦ РАН дает, на наш взгляд, ряд следующих преимуществ. Во-первых, представление документов коллекции в XML-формате допускает их многоцелевое использование, например, как для презентации документов в браузере, так и для их обработки программами анализа данных. Во-вторых, XML-формат стал фактическим стандартом для обмена данными между различными информационными системами, а стандарт метаданных DC – для описания цифровых Интернет - ресурсов. Это должно способствовать интеграции на логическом уровне наших электронных коллекций с другими информационными системами, например, интеграции в Единое научное информационное пространство РАН. В-третьих, технологии и стандарты XML активно развиваются и получили поддержку в программных продуктах ведущих компьютерных фирм. При этом имеется достаточное количество свободно

распространяемого программного обеспечения для работы с данными в XML-формате (например, реализации DOM-модели – http://xml.apache.org , SAX – http://www.saxproject.org и т.п.).

Технология DTD применяется для описания структуры класса документов коллекции, а XML используется для представления класса соответствующих DTD коллекции действительных (valid) документов коллекции. Такой подход позволил нам разработать специальную технологию публикации и сопровождения документов и реализовать с использованием языка Java программный комплекс, обеспечивающий дружественный интерфейс при заполнении полей документов самими специалистами-предметниками, а также автоматическую верификацию целостности документов и формирование соответствующих действительных XML-документов [9,10]. Для представления документов в браузере разработаны специальные PHP-скрипты. С помощью СУБД MySQL на основе атрибутов стандарта DC создана база данных, описывающая общие свойства коллекций ЭБ.

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.

Полученный в ходе разработки и реализации проекта опыт показал следующее. Во-первых, работа по созданию ЭБ научных информационных ресурсов способствовала активизации деятельности сотрудников центра по систематизации и структуризации своих научных информационных ресурсов не только в виде научных публикаций, но и в виде предметных электронных коллекций документов. Во-вторых, выбор стандарта описания метаданных Dublin Core и разработанная технология публикации и сопровождения научных коллекций вполне оправдали наши надежды, связанные с привлечением к этой работе специалистов-предметников. Это подтверждается, в частности, как созданными коллекциями, так и формированием и включением в ЭБ метаописаний созданных ранее в центре электронных информационных ресурсов (баз данных, ГИС-систем и т.д.) и списков ссылок на подобные научные Интернет-ресурсы. В-третьих, разработка программных сервисов осуществляется параллельно с работой специалистов-предметников по созданию научных коллекций. Это позволяет нам учесть их требования и пожелания при разработке программных сервисов ЭБ. В то же время необходимо отметить, что в связи с отсутствием в нашем распоряжении подходящей XML СУБД нам приходится решать и ряд технических вопросов по эффективной реализации необходимых программных сервисов. В частности, в этом плане следует отметить предложенную нами в данной работе реализацию сервиса поиска и сортировки XML-документов в ЭБ КарНЦ РАН.

3 Развитие программных сервисов – задача поиска и сортировки XML-документов.

Вопросы хранения и поиска XML-документов рассмотрены достаточно подробно, например, в работе Марка Грейвса [11]. В данной работе задачу поиска XML-документов в коллекциях ЭБ КарНЦ РАН сформулируем следующим образом. Обозначим через $\mathbf{D=\{d_1,d_2, ..., d_n\}}$ – множество XML-документов коллекции, каждый из которых хранится в отдельном файле **j.xml, j=1, …, n**. Предполагается, что для каждой коллекции задано **m-** элементов описания структуры документа, на основе значений которых будет проводиться отбор искомых документов из данной коллекции. Необходимо разработать процедуру, обеспечивающую поиск искомых документов в коллекциях ЭБ по запросам пользователя, а также их сортировку в алфавитном порядке по названиям описываемых в документах объектов и последовательный вывод списка найденных документов на экран монитора (для удобства – порциями по 25 и менее названий).

Для решения поставленной задачи мы рассмотрели несколько подходов. Например, можно открывать файлы **j.xml, j=1, …, n** по очереди и с помощью DOM-модели (интерфейсы которой используются для доступа к элементам XML-документа) реализовать алгоритм поиска искомых документов, а затем их отсортировать. Недостатки этого решения очевидны: требуется **n**- операций открытия файлов, а с ростом количества документов увеличиваются вычислительные затраты на их сортировку. Для сокращения количества операций открытия файлов можно объединить все файлы **j.xml, j=1, …, n** в один файл всех документов коллекции. Недостатком является то, что при использовании DOM-модели резко возрастают затраты оперативной памяти, связанные с формированием объектной модели этого “большого” XML-документа.

Одним из общеизвестных способов ускорения поиска данных в БД является индексация [9]. Для решения поставленной задачи нами предлагается следующая структура индексного файла коллекции, представленного в виде XML-документа.

```
<indexing>
    <описание_вида  id=”1”  … >
        <искомый_элемент_1>значение_элемента
        _1</искомый_элемент_1>
        <искомый_элемент_2>значение_элемента
        _2</искомый_элемент_2>
        …
        <искомый_элемент_m>значение_элемент
        a_m</искомый_элемент_m>
    </описание_вида>

    <описание_вида  id=”2”  … >
        <искомый_элемент_1>значение_элемента
        _1</искомый_элемент_1>
        <искомый_элемент_2>значение_элемента
        _2</искомый_элемент_2>
        …
        <искомый_элемент_m>значение_элемент
        a_m</искомый_элемент_m>
        </описание_вида>

        …

        <описание_вида  id=”n”  … >
        …
    </описание_вида>
</indexing>
```

Индексный файл для каждой коллекции формируется автоматически, на основе ее документов (**j.xml, j=1,…,n -** файлов) и заданного для каждой коллекции списка элементов описания структуры документов, по которым возможен поиск.. При этом предполагается, что все элементы <описание_вида> индексного файла будут заранее отсортированы по элементу <название_на_русском_языке>, который используется системой для вывода списка документов на экран монитора.

Процедура поиска XML-документов в коллекциях ЭБ с использованием индексного файла и DOM-модели заключается в следующем.

1. Проверяется: есть ли в коллекции документы с датой модификации файла больше, чем дата модификации индексного файла? Если да, то запускается процедура переиндексации и поиск будет осуществлен по “новому” индексному файлу. Если нет, то поиск идет по “старому” файлу – **index.xml**.
2. Из файла **index.xml** создается объектная модель документа с корневым элементом <коллекция>. Поиск осуществляется по всем элементам <описание_вида>, а “внутри” каждого “поддерева” – по “непустым” элементам из заполненной пользователем интерфейсной формы поиска. По результатам поиска формируется список отсортированных искомых документов, который выводится на экран монитора по 25 (или менее) названий.

Предложенная процедура поиска XML-документов имеет, на наш взгляд, ряд следующих преимуществ. Во-первых, поиск осуществляется по заранее отсортированным в файле **index.xml** данным. Во-вторых, использование DOM-модели для доступа к элементам индексного файла занимает меньше оперативной памяти по сравнению с доступом к элементам файла, включающего все документы коллекции. При этом необходимо отметить, что если число элементов, по которым возможен поиск искомых документов, приближается к общему числу элементов, описывающих документ в целом, то индексный файл превращается по сути в файл, включающий все документы коллекции. В-третьих, в данном случае задействована только одна операция чтения файла. Недостатком предложенного подхода является то, что требуется создание отдельного индексного файла и его периодическое обновление.

4 Развитие информационного содержания (контента) ЭБ КарНЦ РАН

В настоящее время развитие информационного содержания ЭБ КарНЦ РАН осуществляется по следующим основным направлениям:

- *Пополнение документами ранее созданных коллекций и создание новых предметных коллекций.* В этом плане пополняются новыми документами три ранее созданные коллекции научных информационных ресурсов: “Аффилофороидные грибы Карелии” (150 документов), “Виртуальная флора Карелии” (99 документов) и “Млекопитающие Карелии” (31 документ). Начаты работы по созданию, на основе разработанной в среде Access базы данных по биотопам Карелии, соответствующей цифровой коллекции документов, которая на первом этапе будет включать свыше 100 описаний биотопов. При этом, каждое описание биотопа будет включать характеристики экотопа и растительности, а также данные об обитающих в нем видах растений и животных, включая электронные фотографии исследуемых объектов, карты-схемы и др. Основой создания предметной коллекции является большой объем научной информации, собранной сотрудниками институтов биологии и леса Карельского научного центра РАН и размещенный в локальных базах данных. Для автоматизации процесса формирования документов коллекции разрабатывается специальное программное обеспечение, предназначенное для поиска нужной информации в локальных базах данных, ее преобразование и включение в соответствующие места XML – документов коллекции.

- *Подключение к поисковому сервису ЭБ информации об электронных научных информационных ресурсах.* Для проведения инвентаризации созданных в центре электронных научных ресурсах разработаны специальные анкеты, в основу которых положены атрибуты стандарта DC. При этом осуществляется сбор информации как о ранее созданных в центре

научных Интернет-ресурсах, так и об электронных научных информационных ресурсах, разработанных с использованием СУБД, ГИС и других систем, с указанием целесообразности создания на их основе соответствующих коллекций в ЭБ КарНЦ РАН. Пополнение этой информацией базы метаданных ЭБ позволяет задействовать поисковый сервис системы для доступа пользователей Интернет в единое электронное информационное пространство центра.

Также проводятся работы по формированию списков ссылок на электронные научные Интернет - ресурсы, соответствующих тематике исследований ученых центра, и их подключению к поисковому сервису ЭБ.

- *Формирование цифровых коллекций электронных версий научных публикаций сотрудников центра.* Здесь, в первую очередь, речь идет о научных публикациях, издаваемых в редакционно-издательском отделе КарНЦ РАН. С этой целью разработан проект соглашения с авторами о их согласии поместить электронную версию своей работы в PDF-формате в коллекции ЭБ. При этом ссылки на электронные публикации размещаются в соответствующих разделах предметного каталога ЭБ.

5 Заключение

В статье описывается текущее состояние работы по созданию и развитию программных сервисов и информационного содержания (контента) электронной библиотеки научных информационных ресурсов КарНЦ РАН – http://dl.krc.karelia.ru. Представлено краткое описание основных моментов разработки и реализации текущей версии системы на основе платформы XML. Рассматривается задача построения эффективной процедуры поиска и сортировки XML-документов в коллекциях ЭБ с использованием возможностей индексации и DOM-модели, а также представлены основные направления работы по развитию информационного содержания (контента) ЭБ КарНЦ РАН.

В настоящее время проводятся работы как по совершенствованию существующих и разработке новых сервисов системы, так и по развитию информационного содержания (контента) ЭБ.

Литература

[1] Бездушный А. Н., Жижченко А. Б., Каленов Н. Е., Кулагин М. В., Серебряков В. А., Бездушный А. А. Предложения по наборам метаданных для научных информационных ресурсов ЕНИП. //Труды шестой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции ", Пущино, 29 сентября-1 октября 2004 г. с. 277-284.

[2] Бездушный А. А., Бездушный А. Н., Нестеренко А. К., Серебряков В. А., Сысоев Т. М. Возможности технологий ИСИР в поддержке Единого Научного Информационного Пространства РАН. //Труды шестой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции ", Пущино, 29 сентября-1 октября 2004 г. с.254-263.

[3] Ю. И. Шокин, В. А. Ламин, А. М. Федотов, В. Б. Барахтин, О. Л. Жижимов, Н. А. Мазов, Б. Н. Пищик, Н. Н. Покровский Распределенная информационная система "Виртуальный музей науки и техники СО РАН". //Труды пятой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции ", Санкт-Петербург, 29-31 октября 2003 г. с.112-116.

[4] Когаловский М.Р. Систематика коллекций информационных ресурсов в электронных библиотеках. //Программирование. № 3, 2000 г. с. 31-52.

[5] Когаловский М.Р. Стандарты XML и электронные библиотеки. Электронные библиотеки - 2003 - Том 6 - Выпуск 2.

[6] В. Вдовицын, А. Сорокин Вопросы формирования и использования электронных научных информационных ресурсов. //Информационные ресурсы России.- 2004 -№ 4, с. 7-12.

[7] Концепция создания Единой информационной системы Российской академии наук. Вторая редакция. http://www.ras.ru/scientificactivity/eis/eisconception.aspx .

[8] С.А. Арнаутов Роль и место виртуальных цифровых библиотек в Интернете. // Труды третьей Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции ", Петрозаводск, 11-13 сентября 2001 г. - Карельский научный центр РАН. с. 20-25.

[9] В.Т. Вдовицын, А.Д. Сорокин Технология публикации и сопровождения документов в коллекциях научных информационных ресурсов электронной библиотеки КарНЦ РАН //Труды пятой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции ", Санкт-Петербург, 29-31 октября 2003 г. с.103-105.

[10] 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.

[11] Грейвс, Марк. Проектирование баз данных на основе XML.: Пер. с англ. - М.: Издательский дом "Вильямс", 2002. - 640 с.

Development of software services and the content of the Karelian Research Centre of the RAS Digital Library

Vladimir T. Vdovitsyn, Anatoly D. Sorokin,
Natalia B. Lugovaya

The paper reports interim results of the activities for the development and advacement of software services and information content of the Digital Library of the Russian Academy Centre scientific information resources (DL KarRC RAS) – http://dl.krc.karelia.ru/ .

♣Работа выполняется при финансовой поддержке РФФИ (грант № 05-07-90077).