ВОПРОСЫ ПОСТРОЕНИЯ ЭЛЕКТРОННОЙ БИБЛИОТЕКИ КАРЕЛЬСКОГО НАУЧНОГО ЦЕНТРА РАН

Вдовицын В.Т., Сорокин А.Д.
Институт прикладных математических исследований Карельского научного центра РАН
185610, г. Петрозаводск, Пушкинская 11
e-mail: vdov@krc.karelia.ru, sorokin@krc.karelia.ru

В данной работе представлено описание основных компонентов разрабатываемой инфраструктуры электронной библиотеки КарНЦ РАН, реализующих возможности: публикации новой коллекции и ее сопровождение; поиска документов коллекции по запросам пользователя и анализа использования информационных ресурсов посетителями сайта.

SOME POINTS OF THE KARELIAN RESEARCH CENTER'S DIGITAL LIBRARY CONSTRUCTING

Vladimir T. Vdovitsyn, Anatoly D. Sorokin
Institute of Applied Mathematical Research of the Karelian Research Center of the Russian Academy of Sciences
11, Pushkinskaya St., Petrozavodsk, Russia, 185610
e-mail: vdov@krc.karelia.ru, sorokin@krc.karelia.ru

This paper concerns the description of the infrastructure of the Karelian Research Center's Digital Library which is still being formed now. The main components of the infrastructure help users to realize the following abilities:

- the one to publish a new collection and to renovate it;
- the one to search for the documents of the collection;
- the one to analyze the usage of the digital library resources.

1. Формирование электронных информационных ресурсов в КарНЦ РАН

В настоящее время в институтах Карельского научного центра РАН (КарНЦ РАН) активно ведутся работы по созданию и развитию электронных информационных ресурсов с использованием Internet-технологий [1]. С 1997 года создан и функционирует официальный сайт центра (http://www.krc.karelia.ru/), сайты институтов (http://biology.krc.karelia.ru/, http://geoserv.krc.karelia.ru/, http://nwpi.karelia.ru/), а также ряд тематических предметно-ориентированных сайтов с интегрированными базами данных, например, в гуманитарной области (http://phonogr.krc.karelia.ru/,

http://toris.krc.karelia.ru/), в биологии (http://biodiv.krc.karelia.ru/) и др. Большинство из этих проектов выполнялись в течение 1997-2001 г.г. при поддержке грантов РФФИ, РГНФ, ФЦП "Интеграция" и др. При этом использовались традиционные Web-технологии и свободно распространяемые СУБД – miniSQL и MySQL с организацией доступа к базам данных по разработанным интерфейсным формам с применением технологии CGI-скриптов.

Увеличение объема и разнообразия создаваемых научных электронных информационных ресурсов в КарНЦ РАН привело к необходимости создания интегрированной информационной системы центра, предназначенной для поддержки фундаментальных исследований и разработок и построенной по единым согласованным стандартам и технологиям. В качестве основного подхода к решению этой проблемы нами рассматривается технология электронных библиотек (Digital Library) [2, 3].

В настоящей время у нас в стране и за рубежом накоплен значительный опыт в создании электронных библиотек (ЭБ). Следует отметить ЭБ РФФИ (http://elibrary.ru/) и СО РАН (http://www-sbras.nsc.ru/win/elbib/), а также ряд зарубежных систем - The Historic Pittsburgh Digital Library (http://www.pitt.edu/~edwardg/histpitt/survey.html), The Alexandria Digital Library (http://www.alexandria.ucsb.edu/) и т.п. Реализация этих и других проектов ЭБ показала практическую значимость и эффективность использования таких информационных систем в первую очередь для поддержки человеческой деятельности в различных областях науки, образования и культуры.

2. Подход к построению электронной библиотеки КарНЦ РАН

На первом этапе работы по созданию ЭБ КарНЦ РАН необходимо решить следующие основные проблемы:

- разработать инфраструктуру и соответствующее системное программное обеспечение, реализующее основные функции ЭБ;
- сформировать коллекции электронных информационных ресурсов о видовом, популяционном и экосистемном биоразнообразии растительного и животного мира Карелии и ее ресурсном потенциале по согласованным стандартам и технологиям.

В дальнейшем планируется расширить состав коллекций за счет привлечения к этой работе специалистов других институтов центра. При этом создаваемая информационная система должна стать частью Единой Информационной Системы РАН (ЕИС РАН, http://www.ras.ru/EIS/EIS_Concept.htm).

Наш подход к построению ЭБ КарНЦ РАН в самом общем виде заключается в следующем. Предполагается, что на головном сайте ЭБ центра будет размещено разработанное программное обеспечение для поддержки

процессов создания, управления, хранения и использования научных коллекций электронных информационных ресурсов. На тематических сайтах Институтов биологии и леса (в перспективе и др. институтов центра) будут размещены предметные научные коллекции. Доступ к коллекциям для сторонних пользователей будет возможен через головной Web-сайт ЭБ КарНЦ РАН.

Структуры документов коллекций разрабатываются на основе представленных специалистами-предметниками паспортов описаний биологических объектов и реализуются с помощью языка XML (eXtensible Markup Language, http://www.w3c.org/XML/overview.html). Для описания общих свойств каждой коллекции используется стандарт Дублинского ядра (DC, Dublin Core, http://purl.org/dc/documents/).

В данной работе представлено описание следующих трех основных компонентов разрабатываемой инфраструктуры ЭБ КарНЦ РАН, реализующих возможности публикации новой коллекции и ее сопровождение, поиска документов коллекции по запросам пользователя и анализа использования информационных ресурсов ЭБ посетителями сайта.

Основными функциями программного обеспечения, предназначенного для поддержки процессов публикации документов в ЭБ КарНЦ РАН и ее сопровождения, являются:

- помощь в заполнении (корректировке) документа коллекции в соответствии с разработанной структурой - DTD (Document Type Definition) языка XML;
- автоматическое формирование корректного XML-документа коллекции и его сохранение на сервере;
- организация на форуме обсуждения экспертами вносимого в коллекцию документа.

Данная программа разрабатывается с использованием агентной технологии (относиться к классу Interface Agent) и реализуется в виде applet на языке Java, для работы на стороне клиента и application, выполняющего определенные действия на сервере.

Основными функциями программы-клиент являются:

- считывание с сервера соответствующего DTD-файла;
- анализ DTD-файла (выделение элементов, атрибутов, комментариев);
- формирование интерфейсной формы, содержащей: графическое отображение структуры элементов; текстовое поле ввода содержимого элемента; текст технической подсказки для элемента; текст смысловой подсказки-расшифровки названия элемента; поля ввода или выпадающие списки значений атрибутов; технические и смысловые подсказки для атрибутов; кнопки перехода к предыдущей и последующей версии элемента (в случае разрешенной повторяемости), а также кнопки отправки готового документа на сервер;

- проверка введенного текста на соответствие техническим требованиям к текущему элементу или атрибуту (тип, повторяемость), замена недопустимых символов (<, >, ` и т.п.) специальными последовательностями (&lt, &gt, &apos и т.п.);
- формирование XML-документа (вставка в текст инструкций обработки, тегов и атрибутов) и отправка его на сервер;
- «выставление» документа на тематическом форуме электронной телеконференции и оповещение об этом по электронной почте экспертов с целью получения от них замечаний и предложений по содержанию документа.

Основной функцией программы-приложения, работающей на стороне сервера, является сохранение сформированного XML-документа в файле документов коллекции.

Процедура поиска информационных ресурсов в ЭБ КарНЦ РАН разбивается на три основных этапа:

- поиск по рубрикатору;
- поиск научной коллекции информационных ресурсов с использованием базы системных метаданных;
- поиск документа в коллекции по заданным критериям отбора данных.

Таким образом, технологию поиска искомого документа в ЭБ КарНЦ РАН можно представить следующим образом. На первом шаге пользователь находит по рубрикатору интересующую его область знаний. Далее, в этой области знаний он выбирает из имеющихся в ЭБ центра коллекций интересующую его коллекцию. Для этого он формирует с помощью интерфейсных форм запрос к базе системных метаданных, которая построена на основе атрибутов стандарта Дублинского ядра. На третьем шаге пользователь формирует запросы на поиск документа в выбранной им коллекции. При этом интерфейсные формы запросов на поиск документа в коллекции реализуются с учетом специфики предметной области и включает наборы шаблонов, которым должен удовлетворять искомый фрагмент документа (примером подобного языка запросов является XML-QL, http://www.w3c.org/TR/NOTE-xml-ql).

В системе предполагается также задействовать процедуру анализа использования информационных ресурсов пользователями ЭБ КарНЦ РАН на основе применения алгоритмов поиска регулярных эпизодов [4]. Задача поиска регулярных эпизодов применительно к анализу посещений сайта может быть сформулирована следующим образом. На основе информации об истории посещений сайта, хранящейся в регистрационном файле Web-сервера, необходимо определить маршруты переходов пользователей по страницам сайта, в которых найти наиболее характерные (часто встречаемые) составляющие их фрагменты. Для решения этой задачи нами разработана и реализована на Java программная система Dminer[5], с помощью

которой на основе анализа регистрационных файлов (лог-файлов) можно получить ответы на ряд вопросов, например, определить тематическую направленность интересов посетителей ЭБ, частоту посещения и др. Эта информация может быть использована, например, для повышения эффективности работы программного обеспечения ЭБ КарНЦ РАН.

Работа выполняется при поддержке РФФИ (грант № 02-07-90204).

Литература

1. Сорокин А.Д., Вдовицын В.Т., Луговая Н.Б.. Создание и развитие электронных информационных ресурсов в КарНЦ РАН. Сб. докл. Второй Всеросс. Науч. Конф. "Электронные библиотеки: перспективные методы и технологии, электронные коллекции". Протвино, 26-28 сентября 2000 г., с.3-5.
2. Когаловский М.Р.. Систематика коллекций информационных ресурсов в электронных библиотеках. // Программирование. № 3. 2000 г., с. 31-52.
3. А.Н. Бездушный, А.Б. Жижченко, М.В. Кулагин, В.А. Серебряков. Интегрированная система информационных ресурсов РАН и технология разработки цифровых библиотек. // Программирование № 4. 2000 г., с. 3-14.
4. Spiliopoulou M., Faulstich L.C.. WUM: A Tool for Web Utilization Analysis. // Proc. Of EDBT WebDB'98. Valencia: Springer Verlag, 1999, p.184-203.
5. С.А. Бедорев, Ю.В.Чуйко. Применение алгоритма поиска регулярных эпизодов для анализа посещений Web-сайта. // Методы математического моделирования и информационные технологии. Труды ИПМИ КарНЦ РАН, вып.3., 2002, с.153-169.