Интернет архиватор для библиотек (коробочный вариант)

© Шварцман М.Е.

Российская государственная библиотека
shvarl@rsl.ru

Аннотация

Этот документ описывает опыт создания программного обеспечения для скачивания Интернет ресурсов в библиотеках. Вниманию предлагается готовый продукт, который с успехом применяется в РГБ для создания архива онлайновых журналов.

1 История разработки

В Российской государственной библиотеке с 2004 года идет работа по созданию портала российских научных журналов, размещенных в Интернете, проект **«Создание архива российских научно-технических полнотекстовых журналов, опубликованных в Интернет», получивший поддержку РФФИ, грант 04-07-90056-в**.

В ходе работы испытывались разные варианты программного обеспечения для скачивания журналов и организации поиска в полученном массиве информации. Изучался опыт использования DSpace, большое время было уделено испытанию программного обеспечению Greenstone, как средства организации электронной библиотеки для хранения скачанных журналов, были внесены доработки в его исходный код и даже был сделан опытный образец. однако впоследствии от этой идеи пришлось отказаться. Хотя наш опыт использования Greenstone и показал, что система эта довольно удобная и простая в освоении. Коллектив разработчиков довольно оперативно отвечает на все вопросы. Сообщество пользователей многочисленно многоопытно и доброжелательно. Мы по прежнему используем Greenstone в РГБ, но для других задач, там где нужно осуществить поиск по элементам библиографического описания, сформировать электронную библиотеку на компакт диске для распространения т.п. Дело в том, первоначально мы надеялись построить алгоритм автоматического распознавания автора и заглавия статьи в скачанных журналах. Такая библиографическая информация позволила бы использовать многие свойства Greenstone, Однако нам так и не удалось построить надежно работающий механизм распознавания и поэтому мы решили скачивать журналы, как не структурированный текст. В таком тексте все преимущества Greenstone пропадают и для полнотекстового поиска нам оказалось проще использовать бесплатное программное обеспечение mnogosearch http://www.mnogosearch.org, увязанное в общий пользовательский web интерфейс.

2 Что удалось сделать

В настоящее время работа близится к завершению, и нам удалось создать работоспособное и, что немаловажно, переносимое решение, которое, как нам кажется, может быть использовано в любой заинтересованной организации или любым исследователем для формирования своей собственной электронной библиотеки путем скачивания из Интернета необходимых ресурсов.

Мы поставили перед собой задачу сделать программное обеспечение достаточно универсальное, настраиваемое, несложное в установке, распространяемое бесплатно, с исходными кодами, с помощью которого можно создать архив ресурсов Интернет.

Что же в итоге у нас получилось.

Наша система состоит их трех модулей:

Каталог ресурсов

Архив ресурсов

Полнотекстовый поиск в архиве.

Все модули могут функционировать самостоятельно, могут быть установлены на отдельные компьютеры и взаимодействовать между собой по сети.

Каталог ресурсов – это база данных, реализованная на PostgreSQL (выбор СУБД во многом определялся требованием свободного распространения). Обращение к БД производится двумя путями. Для ввода библиографического описания (БО) и редактирования его, а также для поиска по

элементам БО разработан WEB интерфейс, позволяющий вести работу с любого компьютера в Интернет. Форматом описания ресурсов выбран Dublin Core Metadata Element Set (DC MES). Мы используем все поля этого формата и часть рекомендуемых квалификаторов.

Для проведения административных операций типа переиндексирования базы, ввода новых пользователей и экспорта-импорта записей была разработана специальная клиентская программа, работающая под Windows и выполняющая все эти функции. Кроме БО в базе данных содержится информация о параметрах для скачивания полных текстов статей в нашем случае или просто отдельных страниц сайтов в общем случае. Администратор задает время, через которое нужно проверить сайт на обновление и количество копий, которые нужно хранить. Модуль скачивания основан на свободно распространяемой программе WGET, работающей под всеми версиями UNIX или под Windows с использованием библиотек Cgwin. Эта программа была доработана, и к ней был дописан отдельный модуль на PHP.

Как выглядит архитектура хранилища.

В настоящий момент одновременно хранится три копии сайтов. В начальный момент работы программа скачивает полную версию сайта в соответствии с заданным URL (если журнал распределен по двум или более серверам, то указываются несколько URL). Эта копия будет храниться в файловой системе для сравнения при последующем скачивании. Для полнотекстового поиска эта копия проходит обработку для удаления рекламных блоков и прочей непрофильной информации. Администратор задает список регулярных выражений, в соответствии с которыми удаляется ненужная информация. Через заданный для этого сайта интервал времени происходит повторное скачивание. При этом происходит сравнение скачиваемых страниц с копией, оставленной для скачивания. Если различия (в байтах) больше 0, но меньше 20%, от скачиваемого объема, то происходит обычная операция замены старой копии новой, результат прошлого скачивания становится второй копией. Если отличия больше чем 20%, то информация об этом доставляется администратору для принятия решения о допустимости замены копии. Возможны случаи, изменения содержимого сайта без изменения URL малоценной информацией, и при этом необходимо избежать затирания содержимого сохраненного сайта. Сейчас в нашем архиве около 800 журналов, общий объем архива составляет 300 гб.

Ознакомится с работающей системой можно по адресу http://j.rsl.ru/.

Опыт архивирования журналов показал, что возникающие проблемы можно поделить на три класса

1. Технические проблемы, связанные с выявлением типичных ошибок в использовании HTML, возникновением новых методов организации сайтов и в связи с этим необходимости доработки программы.
2. Организационные проблемы, связанные с необходимостью организации специальной службы, отслеживающей появление новых журналов и занесение их в базу, регулярное администрирование системы и своевременное добавление жестких дисков.
3. Юридические проблемы, связанные с невозможностью на законных основаниях скачивать онлайновые журналы и публиковать в виде архива.

Если первые две проблемы можно решить при наличии финансирования, то третья пока не решается.

Литература:

1. Шварцман М.Е. Архив российских научных онлайновых журналов.//Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Седьмой Всероссийской научной конференции (RCDL'2005). Ярославль, 4-6 октября 2005 г.

2. Нудель С. А. К вопросу об архивировании электронных ресурсов [электронный документ] //EVA 2004 Москва

Internet Archiver for libraries in box

Shvartsman Mikhail

This document describes experience of creation of the software for harvesting the Internet resources for libraries. We suggest product is applied in Russian State Library to creation of archive of on-line magazines