XXIX Неделя науки СПбГТУ. Материалы межвузовской научной конференции. Ч.V: С.43-44, 2001.

УДК 681.3.61

А.В. Команцев (5 курс, каф. АиВТ), Н.В. Соколова, к.т.н. доц.

ПРЕДСТАВЛЕНИЕ СПИСКА WEB-ПУБЛИКАЦИЙ В ВИДЕ ЭЛЕКТРОННОГО КАТАЛОГА

В настоящее время сеть Internet является одним из основных хранилищ информации, и актуальной задачей является представление эффективных средств поиска электронных ресурсов в Internet. Существует два основных подхода к созданию средств поиска информации: по полному тексту (Search Engine) и по описанию документа. Представляемое исследование связано с созданием средств автоматизированной обработки электронного ресурса, позволяющей единообразно, стандартным способом, генерировать описания ресурсов и таким образом предоставлять информацию для поиска документа в Интернет.

Таким образом, задача исследования - создание программного обеспечения для автоматизации описания электронных ресурсов, создание каталога и разработка среды для предоставления средств эффективного поиска информации в глобальной сети с помощью каталога. Искомая система должна обеспечивать поиск ресурсов любого типа, не ограничиваясь рамками какой-либо одной предметной области.

Для представления пользователю информации об электронных документах необходимо собрать и занести в базу данных описания каждого из собираемых ресурсов. Для удобства систематизации и для облегчения поиска необходимо, чтобы все описания соответствовали какому-либо из принятых стандартов. В проекте принято решение использовать рекомендации DC (Dublin Core, http://purl.org/dc), как общепринятый и наиболее перспективный стандарт описания ресурсов. На этом стандарте базируются синтаксические основы для Web-ориентированных метаданных, которые получили название RDF (Resource Description Framework, http://www.w3.org/RDF/). Описание ресурса по рекомендациям DC состоит из 15 полей, в которых описываются основные характеристики ресурса.

Поля DC могут быть реализованы различными средствами. В том числе известными языками разметки HTML и XML. В исследовании проведено их сравнение, в результате чего был выбран формат XML, поскольку он предоставляет удобные средства структуризации.

Исходя из постановки задачи, в архитектуре разрабатываемой программы можно выделить четыре основных блока:

- Разбивка аннотированного списка публикаций на отдельные ссылки. При занесении списка публикаций в базу данных необходимо по заданному URL-адресу перечня проанализировать все гиперссылки.
- Получение DC-описания ресурса. Ресурсом может быть как электронный документ, находящийся по определенному библиотекарем URL-адресу, так и один из ресурсов, полученных в результате анализа перечня на предыдущем шаге. Для каждой ссылки надо составить максимально полное описание, соответствующее принятым в данном проекте стандартам.
- Занесение полученного описания в базу данных. Полученные описания после просмотра их библиотекарем и возможной коррекции необходимо занести в базу данных, тем самым сделав эти описания доступными для поиска.
- Поиск записи по базе данных по ключевым словам. Имеющиеся в базе данных описания должны быть предоставлены пользователю каталога по запросу в соответствии со сформулированным пользователем запросом.

Запросы от пользователя или промежуточные данные между блоками программы могут передаваться системе двумя способами: при помощи командной строки и посредством Web-интерфейса.

Проект был реализован на языке PERL. В результате исследования было создано программное обеспечение, полностью удовлетворяющее заданным требованиям. Результаты тестирования показали, что разработанный проект работоспособен и может применяться для организации распределенных электронных библиотек и каталогизации веб-ресурсов. Но, несмотря на то, что программное обеспечение является законченным и самостоятельным продуктом, ведется доработка и усовершенствование некоторых его частей.

В качестве основных направлений развития программной системы выделены следующие:

- Расширение поисковых возможностей системы. При определении содержания документа через ключевые слова (в общем случае, нестандартные) может снизиться релевантность поиска. По этой причине целесообразным является создание средств навигации по предметным рубрикам, которые являются стандартными на национальном уровне. В России приняты предметные рубрики УДК и некоторые другие по более узким предметным областям.
- Поддержка актуальности базы данных и целостности ссылок. Размещение и содержание ресурса в Интернет не является постоянным, поэтому необходимо осуществлять постоянный контроль достоверности описания ресурсов, находящихся в электронной библиотеке. Планируется создание программы-робота, которая периодически в часы минимальной загрузки системы запросами пользователей сверяла бы содержание каждого ресурса с его описанием в базе.

Размещение созданного проекта в сети и активная работа с ним в реальных условиях (а не только тестовые испытания) помогут точнее определить первоочередные задачи в дальнейшем развитии системы. Создаваемый компонент является модулем Электронной Библиотеки СПбГТУ, и в дальнейшем будет использован при создании образовательного портала.