ПОГОРЕЛКО Константин Павлович – *кандидат технических наук, зав. библиотекой Математического института им. В.А. Стеклова (отдел БЕН РАН).*

ПРОГРАММНЫЕ СРЕДСТВА ДЛЯ СОЗДАНИЯ ЭЛЕКТРОННЫХ КНИЖНЫХ КОЛЛЕКЦИЙ

Описываемый программный комплекс предназначен для электронной публикации полнотекстовых документов, которые представлены в виде набора изображений страниц первоисточника. Комплекс обеспечивает поддержку всей технологической цепочки от «грязного» изображения, полученного со сканера, до представления готового документа в интернете.

Основной идеей предлагаемого подхода является предоставление возможности пользователю постраничного просмотра полнотекстового документа и, при наличии полномочий, получения интересующих его частей документа или документа целиком в виде единого файла. Пользователь, найдя интересующую его публикацию, имеет возможность просмотреть отдельные страницы и решить, какие части документа ему нужны. Этот подход позволяет, с одной стороны, существенно сократить трафик для документов большого объема, какими являются книжные коллекции. Пользователю пересылаются только затребованные им страницы и с разрешением, соответствующим разрешению экрана. С другой стороны, этот подход позволяет сохранить контроль над документом со стороны электронной библиотеки. Пользователь может быть ограничен в размерах затребованной информации.

При предоставлении возможности постраничного просмотра для удобства пользователя должна обеспечиваться возможность быстрого нахождения пользователем интересующих его разделов документа. Для обеспечения этой возможности документ снабжается электронным оглавлением с возможной иерархической структурой, позволяющим осуществлять быструю навигацию в пределах документа.

Средства позволяют организовывать и обслуживать ссылки не только ко всему документу целиком, но и к отдельным его частям. Это важно при электронной публикации сборников и конволют. Для этих изданий можно обеспечить доступ пользователю к отдельным публикациям, входящим в документ.

При реализации программного комплекса предполагались следующие стандарты на сканирование документов:

- Отдельное изображение для каждой страницы.
- Разрешение 600 dpi.
- Формат tiff, сжатие CCITT G4 (ITU-T T6) для черно-белых изображений.
- Формат tiff, 8 бит на канал, без сжатия для серых и цветных изображений.

Программный комплекс реализован на языке C# в среде Microsoft .net 2.0 и функционирует на операционных системах Windows. Серверные части, осуществляющие преобразования графических файлов «на лету», в целях обеспечения эффективности реализованы на C++ с использованием технологии ISAPI. Функционально программный комплекс состоит из трех частей:

- Средства подготовки изображений;
- Средства загрузки изображений на сервер и ввода структуры документа;
- Средства представления информации пользователю.

Для подготовки изображений к публикации используется программа редактирования изображений в формате tiff. Основное назначение данной программы – обеспечить максимально эффективное устранение дефектов, возникающих при сканировании. Данная программа позволяет осуществлять обрезку изображений до выбранного размера, очищать поля изображения, устранять ряд дефектов сканирования. Для устранения дефектов сканирования в программу заложено два алгоритма. Первый позволяет устранить так называемый «белый шум» – вкрапление белых пикселей в начертания букв, который возникает при сканировании с недостаточными уровнями яркости и контрастности. Этот алгоритм также сглаживает очертания букв. Второй алгоритм предназначен для устранения «серого шума» - скоплений черных пикселей, возникающих на белом фоне при плохой бумаге источника и избыточных уровнях яркости и контрастности. Размер удаляемых пятен может задаваться оператором при настройке программы.

Программные средства загрузки изображений предназначены для загрузки изображений документа на сервер и ввода и редактирования электронного оглавления. Средства реализованы в виде клиентской и серверной частей. Клиентская часть обеспечивает эффективный диалог оператору системы, а серверная часть обеспечивает необходимые сервисы для клиентской части. Взаимодействие клиентской и серверной частей обеспечивается на внутреннем протоколе системы, реализованном на базе протокола http. Подобное решение обеспечивает, с одной стороны, удобство и эффективность работы оператора системы и, с другой стороны, высокую защищенность сервера.

Каждый элемент электронного оглавления снабжается ссылками на начальный и конечный номер изображения соответствующего раздела. Начальный номер используется для навигации по документу. Для выдачи раздела в виде единого файла используются оба номера. Кроме того, элемент оглавления может

содержать отметку, указывающую на то, что он является внешней ссылкой. Для ускорения ввода электронного оглавления можно использовать заранее заготовленный текстовый файл. Подобный файл получается путем графического распознавания существующего книжного оглавления и его дальнейшего редактирования. Для проверки соответствия ссылок элементов оглавления на соответствующие изображения страниц система позволяет просматривать необходимые страницы.

Презентационная часть комплекса представляет собой web-сайт, обеспечивающий представление пользователю документа, номер которого содержится в запросе. Экран, предоставляемый пользователю, состоит из трех окон: окно навигации, содержащее оглавление, функциональное окно, содержащее элементы, обеспечивающие необходимую функциональность, и окно изображения страницы. Для обеспечения возможности динамического перераспределения пространства экрана между окнами размещение окон осуществлено набором фреймов. Презентационная часть позволяет менять масштаб предоставляемого пользователю изображения. Пересчет изображения из исходного высокого разрешения в 600 dpi в разрешение, отображаемое экраном, происходит на сервере. В системе реализован алгоритм преобразования, позволяющий сохранить читаемость текста, которая часто нарушается при преобразовании черно-белых изображений, осуществляемых большинством известных браузеров. По желанию пользователя масштаб изображения может быть изменен. Существует возможность поворота выдаваемого изображения. По запросу пользователя ему может быть предоставлена какая-либо часть документа или документ целиком в виде единого файла в формате pdf.

Описываемый комплекс применяется на практике в проектах «Электронная библиотека Математического института им. В.А. Стеклова» и «Научное наследие РАН».