Компьютерная система для создания и поддержки электронных коллекций старинных книг

В. С. Южиков
Казанский государственный университет, Россия

In this article the system for creation and support of the digital collections of ancient books is described. The process of creation consists of several stages. At first the user chooses images for the future collection and also creates necessary structure of representation. At the next stage the user choses images that are processed in an automatic or semi-automatic mode for elimination of spots and non-regular colors of paper and ink. Further the user adds and edits metadata for all images. Then on the basis of this data the web-collection is formed.

1. Введение

При создании электронных библиотек необходимо учитывать специфику публикуемого контента, который влияет на структуру хранения информации, ее поиск и отображение. Наиболее распространены электронные коллекции, содержащие преимущественно информацию текстового характера с незначительным включением иллюстраций и прочего медиа-контента. В случае же электронных коллекций старопечатного текста и рукописей основной контент составляют оцифрованные изображения страниц. Это делается по нескольким причинам. Во-первых, получить текстовый вариант страниц возможно лишь для книг XVIII–XIX веков. Для более ранних изданий существующие OCR программы типа FineReader'а плохо справляются с распознаванием символов из-за наличия сильных дефектов и помех, а также из-за отсутствия соответствующих словарей, играющих немаловажную роль в качественном распознавании [Соловьев 2003]. Во-вторых, большой интерес может представлять само изображение страницы с содержащимися на нем пятнами, пометками, рисунками; часто это дает дополнительную информацию об авторе, а также об обстоятельствах написания и хранения книги.

Еще одним важным моментом, который следует учитывать при создании коллекции старинных книг, является относительно большой размер исходных сканированных изображений. Даже при использовании форматов сжатия типа JPEG и GIF изображения занимают порядка 100–500 килобайт, что приводит к существенному замедлению работы конечного пользователя с материалами библиотеки. Это особенно заметно при низкой пропускной способности канала связи (модемное подключение). Данную проблему можно решить следующим образом: необходимо, чтобы наряду с исходными изображениями в полном разрешении присутствовали также и изображения для предварительного просмотра, имеющие как можно меньший объем, но сохраняющие при этом «читабельность» текста. Это может быть достигнуто, например, упрощением или удалением фона, который несет меньшую значимость для восприятия текста, а также сокращением информации о цвете элементов страницы. Иначе говоря, необходимо, чтобы изображение каждой страницы было представлено в нескольких видах в зависимости от потребностей пользователя, это позволит сделать работу с библиотекой более эффективной.

2. Анализ задачи

Для создания электронной коллекции, содержащей оцифрованные изображения страниц старинных книг, необходимо разрешить ряд проблем.

1. Создание группы изображений из каждого исходного изображения с разным разрешением.

2. Обработка всех изображений полученной группы в соответствии с их предназначением. Сюда входит:

– устранение пятен, помех, неоднородностей цвета бумаги, просвечивание надписей с обратной стороны листа;

– удаление или замена фона;

– поворот изображения при неточной ориентации страницы во время оцифровки;

– разметка страницы.

3. Сбор и редактирование метаданных, описывающих изображение.

4. Построение библиотеки и ее размещение на сервере.

Для устранения дефектов исходных изображений, которые заметно затрудняют их использование, можно использовать универ-

сальные графические редакторы, например, Adobe Photoshop. Но это занимает много времени и требует соответствующей квалификации оператора, что неприемлемо для обработки сотен и тысяч изображений, входящих в коллекцию. Поэтому желательно иметь специализированную систему, в которой были бы реализованы все необходимые функции с возможностью как полностью автоматической работы, так и с поддержкой ручной коррекции процесса обработки. В основном в публикациях встречаются описания работ, начатых в этом направлении, а также возможные подходы для решения отдельных задач реставрации [Баженов и др. 2001; Грузман и др. 2000; Масевич и др. 2001], поэтому было принято решение о разработке и реализации собственного модуля обработки изображений.

Для создания и ведения электронных библиотек разработано достаточно много систем. К наиболее известным относятся программы GreenStone и D-Space, бесплатно представленные в сети Интернет. Но они ориентированы преимущественно на текстовое наполнение библиотеки с небольшим количеством рисунков в тексте, что плохо подходит для коллекции старопечатных текстов.

Кроме того, желательно, чтобы модуль обработки и модуль создания библиотеки были функционально объединены в единую систему, что позволит более эффективно использовать все средства, предоставляемые такой средой.

Реализация подобного функционала, например, на основе программы GreenStone, требует существенного изменения и дополнения ядра, что является очень сложной задачей. Поэтому реализация такой системы является оправданной как с точки зрения затрат, так и с точки зрения получаемого результата.

3. Описание разработанной системы

3.1. Выбор исходных файлов

На этом этапе оператор должен выбрать файлы изображений, необходимые для построения библиотеки, а также создать иерархическую структуру в виде дерева, которая будет отражать будущую структуру библиотеки. Это делается с помощью перетаскивания нужных файлов из левой части панели в правую. Этот процесс показан на рис. 1.

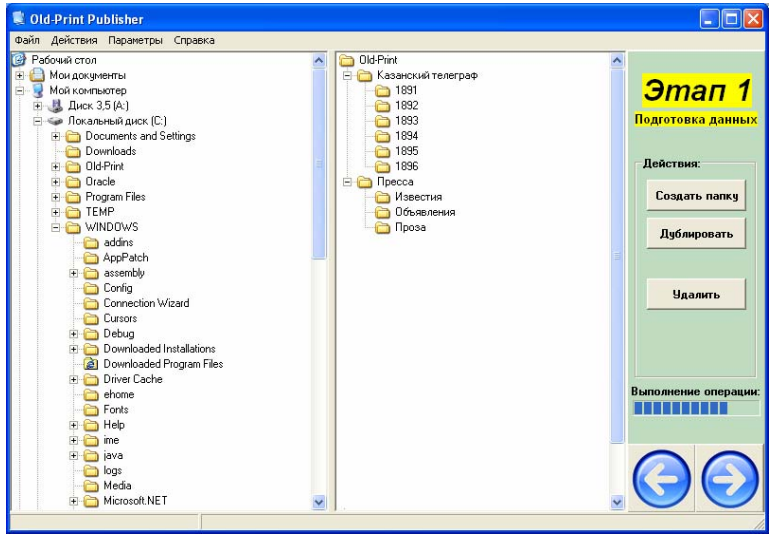

Рис.1. Выбор исходных файлов изображений

3.2. Обработка и реставрация изображений

Для размещения изображений в электронной библиотеке необходимо, как уже описывалось выше, создать определенный набор изображений из каждого исходного изображения. Оператор может выбрать необходимые виды из следующего списка:

1. Изображение, очищенное от технических шумов и помех, но сохраняющее пятна, кляксы и другие дефекты, имеющиеся в оригинале (для проведения различных исторических исследований, изучение условий хранения и т. д.). Разрешение — полное.

2. То же, что и предыдущий вид, но с уменьшенным разрешением, достаточным для чтения текста.

3. Изображение, очищенное от крупных и средних пятен, неравномерностей цвета, проступаний надписей с обратной стороны листа и других дефектов, но с сохранением фактуры и мелких деталей фона. Разрешение — уменьшенное, достаточное для чтения текста.

4. Изображение с удаленным фоном, но с сохраненным цветом букв. Оно занимает значительно меньше места и используется пользователем при недостаточной скорости канала связи. Разрешение — уменьшенное, достаточное для чтения текста.

5. Пиктограмма — для отображения в списке доступных в библиотеке страниц (предпросмотр).

При необходимости оператор может вручную скорректировать параметры обработки, если результаты работы алгоритмов в автоматическом режиме не дали хорошего результата. Внешний вид модуля приведен на рис. 2.

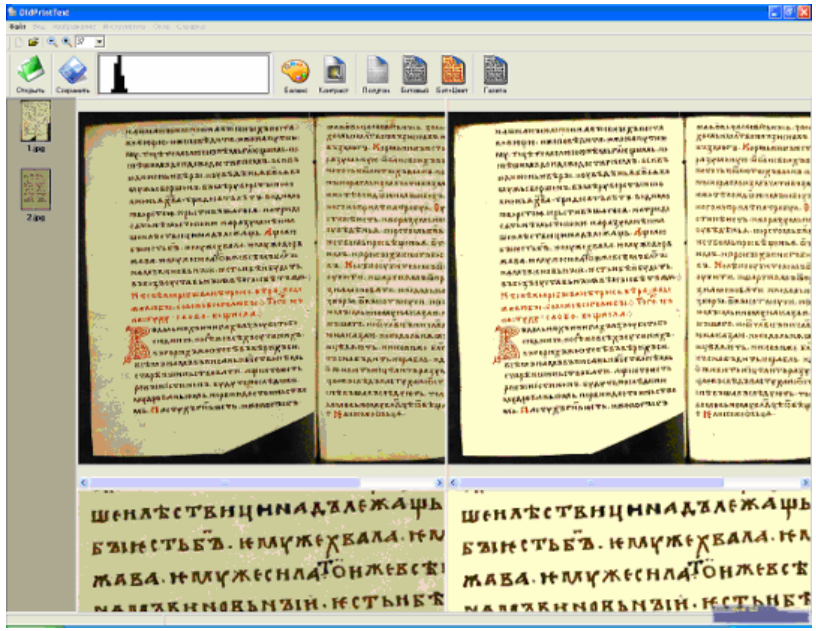

Рис. 2. Модуль обработки изображений

Примеры работы алгоритмов модуля обработки и реставрации изображений приведены на рис. 3–5:

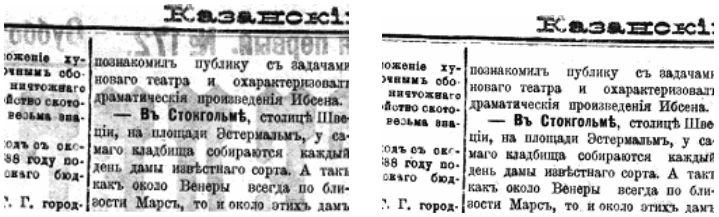

Рис. 3. Устранение проступания надписей с обратной стороны

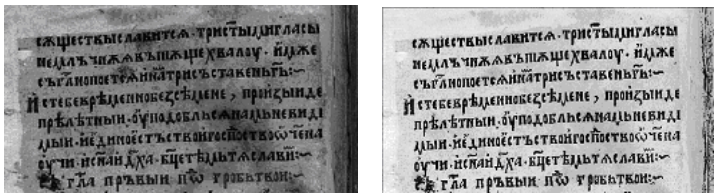

Рис. 4. Выравнивание яркости и контраста

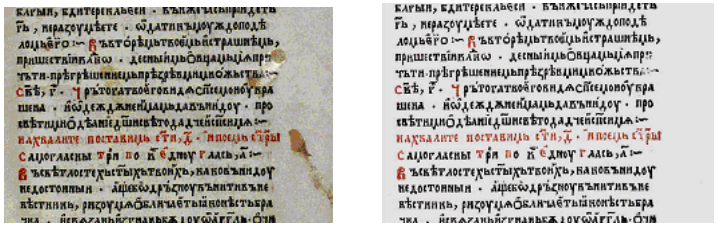

Рис. 5. Удаление пятен

3.3. Редактирование и добавление метаданных

Результатом предыдущих этапов является древовидная структура из файлов изображений, а также набор метаданных, полученных на отдельных стадиях обработки (анализ заголовка файла, результат сегментации, тип обработки). На этом этапе оператор просматривает и редактирует полученные метаданные для каждой группы изображений. В качестве базового был выбран набор метаданных Dublin Core, являющийся общепризнанным базовым стандартом описания.

3.4. Редактирование параметров проекта библиотеки

На этом этапе оператор редактирует параметры проекта библиотеки, такие, как:

– название,
– информация о создателе библиотеки,
– типы поиска,
– внешний вид и другие параметры.

3.5. Построение библиотеки

После того как все необходимые данные собраны и подготовлены, можно перейти к самой процедуре построения библиотеки. Сначала необходимо выбрать тип построения: библиотека для публичного доступа на базе web-сервера либо локальная библиотека (например, для распространения на CD-ROM дисках). Далее

запускается процесс построения, по завершении которого оператор может просмотреть полученный результат и, при необходимости, вернуться на более ранние этапы для корректировки данных. Сформированный проект можно экспортировать в формат электронной библиотеки GreenStone. Также возможен и импорт данных из этого формата.

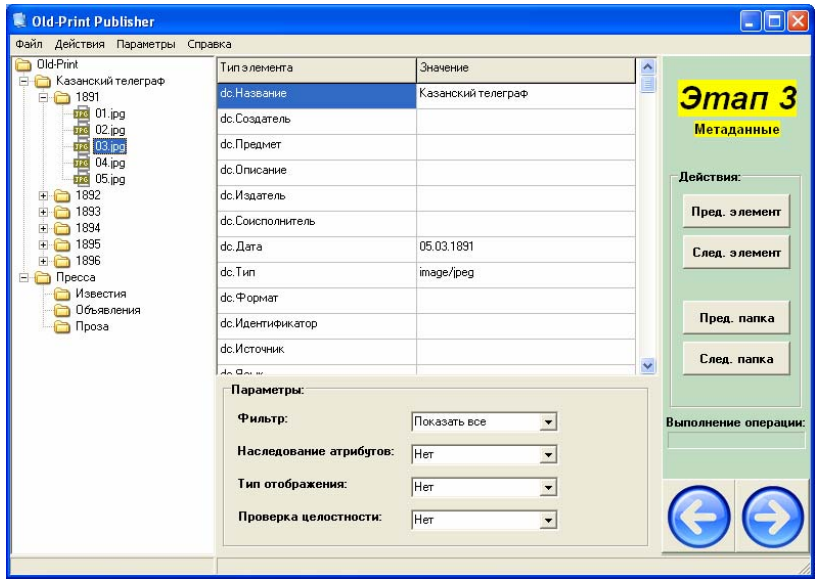

Рис. 6. Модуль сборки метаданных

Заключение

В настоящее время эта система используется в библиотеке Казанского Государственного Университета, а также в Национальной библиотеке им. А. Навои (Республика Узбекистан, Ташкент) при создании электронной коллекции старопечатных изданий.

Описанная система разработана с использованием среды Borland Delphi 7 и языка PHP 5, предназначена для работы на платформе Win9x, Win2k.

Список литературы

Соловьев 2003 — *Соловьев, В. Д.* Электронная коллекция древних книг и рукописей: Исследования по информатике. — Вып. 4. — Казань : ИПИ АН РТ, 2003. — С. 21–26.

Баженов и др. 2001 — *Баженов, С. Р.* Создание цифровых коллекций редких книг и рукописей из сибирских хранилищ / С. Р. Баженов, В. Н. Алексеев, А. Ю. Бородихин, Е. И. Дергачева-Скоп, А. В. Шабанов // Новые технологии в информац. обесп. науки : тр. конф. — М. : Биоинформсервис, 2001. — С. 146–148.

Грузман и др. 2000 — *Грузман, И. С.* Цифровая обработка изображений в информационных системах : учеб. пособие / И. С. Грузман, В. С. Киричук, В. П. Косых, Г. И. Перетягин, А. А. Спектор. — Новосибирск : Изд-во НГТУ, 2000. — С. 69–73.

Масевич и др. 2001 — *Масевич, А. Ц.* К созданию электронных коллекций старопечатных книг в библиотеке Российской академии наук: на примере работы над двумя проектами / А. Ц. Масевич, Е. А. Савельев, А. К. Багажков // Новые технологии в информац. обесп. науки : тр. конф. — М. : Биоинформсервис, 2001. — С.132–140.