Система автоматической обработки сканов книжных фондов для создания электронных коллекций. Опыт сотрудничества с Российской государственной библиотекой

Computerized system for processing book scanned copies to build digital collections. Collaboration with the Russian State Library

Н. С. Дикий
ООО «АЛАНИС Софтвер»,
Новосибирск, Россия

Nikolay Dikiy
ALANIS Software,
Novosibirsk, Russia

В докладе рассматриваются используемые в российских библиотеках практики обработки книжных сканов при оцифровке фондов, а также проблемы, связанные с низкой автоматизацией процессов. В части предлагаемого решения описанных проблем, доклад представляет программное решение массовой обработки книжных сканов компании «АЛАНИС Софтвер» в тесном сотрудничестве с Российской государственной библиотекой.

The technologies of processing book scanned copies during collection digitization used in Russian libraries, and the problems of insufficient computerization are discussed. To solve these problems, the author proposes the software solution of mass book scanned copies processing developed by ALANIS Software Company in close collaboration with the Russian State Library.

Введение

В прошлом году наша компания впервые приняла участие в конференции с докладом о наших разработках в области оцифровки газет и автоматической сборки статей из результатов распознавания. Мы были очень рады интересу, который проявило библиотечное сообщество России к данной теме и нашим разработкам.

Помимо прочего, публикация нашего доклада на сайте ГПНТБ уже принесла свои плоды: ALANIS Software разрабатывает продукт по мониторингу российских печатных СМИ в сотрудничестве с одной из профильных государственных структур.

Мы особенно рады, что результатом нашего участия в конференции в 2015 году стало начало сотрудничество с Российской Государственной Библиотекой.

Актуальность

Прежде чем перейти непосредственно к описанию решения, которое создано нами для РГБ, я бы хотел остановиться на понимании термина «оцифровка» в применении к деятельности библиотек, чтобы объяснить, какую именно проблему мы помогаем решить.

Оцифровка фондов не самоцель, как мы все понимаем. Результатом оцифровки должен являться электронный информационный ресурс, электронный фонд библиотеки. Поэтому качественная оцифровка фондов в современном понимании включает в себя множество промежуточных, часто ресурсоемких операций, необходимых для создания конечного продукта, полнофункционального электронного источника, встроенного в электронную инфраструктуру библиотеки:

- Сканирование
- Обработка сканов
- Распознавание (где применимо)
- Структурирование (на основе результатов распознавания)
- Добавление мета информации (создание электронной карточки источника или привязка к существующей карточке)
- Создание веб портала для онлайн доступа к электронному ресурсу, интеграция с существующими порталами и т.д.

Пропуск одного из промежуточных этапов оцифровки или его некачественная реализация могут привести к полному провалу в создании электронного ресурса или к низкому качеству конечного продукта. Учитывая массовый характер мероприятий по созданию электронных ресурсов библиотек, каждая недоработка или ошибка многократно увеличивает затраты на создание качественного конечного продукта. При этом, ошибки, допущенные на начальных этапах, дополнительно увеличивают затраты на их исправление. В этом смысле, качество исходного скана имеет определяющее значение для успеха всего процесса оцифровки. Именно эту задачу необходимо было решить в рамках сотрудничества с Российской Государственной Библиотекой.

При общении с управлением информационных ресурсов библиотеки обозначился круг проблем, касающихся обработки уже отсканированных фондов РГБ, а это – миллионы сканов книжных страниц, ожидающих «своего часа»: дальнейшей обработки, которая позволит открыть доступ читателям к данным ресурсам в электронном формате.

Думаю, что сканирование фондов осуществляется практически во всех государственных библиотечных учреждениях, поэтому наш опыт, полученный в совместной работе с РГБ и воплощенный в программном продукте, окажется полезным и в других библиотеках, поскольку перед всеми учреждениями стоят схожие задачи по переводу фондов в электронный формат.

Сканирование начинается с момента приобретения дорогостоящего сканирующего оборудования, а именно, планетарных книжных сканеров. Однако, высокая стоимость сканера не означает автоматического получения сканов, готовых для публикации в электронных коллекциях. В данном смысле, ожидания часто расходятся с реальностью. Сканеры поставляются без программного обеспечения по пост обработке, либо специализированное ПО сопоставимо по стоимости со сканером, имеет ограничения по сроку использования и от него отказываются при покупке сканера. Если же ПО так или иначе предоставлено в распоряжение библиотеки, оно может не иметь необходимых функций или не обеспечивает необходимую производительность, так как количество приобретаемых сканеров обычно измеряется штуками, и производительность поставляемого с ними ПО обработки не превышает возможности сканера, а часто просто не предназначено для обработки «посторонних» файлов. Поэтому сканирование в библиотеках часто проводится как отдельная операция, без постобработки; таким образом, накапливается огромное количество сканов «полуфабрикатов», которые требуют выравнивания, обрезки, а также других оптимизирующих операций, прежде чем из них можно будет создать электронные ресурсы.

Дополнительным «толчком» для начала работ по созданию продукта, способного решить обозначенные задачи стала декларированная государством политика импортозамещения, и, в частности, статья 12.1 Федерального закона от 27 июля 2006 г. № 149-ФЗ «Об информации, информационных технологиях и о защите информации» (в редакции Федерального закона от 29 июня 2015 г. № 188-ФЗ) и постановление Правительства Российской Федерации от 16 ноября 2015 г. №1236 «Об установлении запрета на допуск программного обеспечения, происходящего из иностранных государств, для целей осуществления закупок для обеспечения государственных и муниципальных нужд».

Задачи

Какого рода продукт необходим в описанной ситуации?

В общем виде требования можно сформулировать следующим образом:

В функциональном смысле продукт должен обеспечивать автоматическую обработку набора «сырых» отсканированных изображений одного источника (например, одной книги) таким образом, чтобы на выходе получилась книга в PDF формате без каких-либо артефактов сканирования с единообразными характеристиками страниц на всем протяжении источника. Комплементарным форматом на выходе должны быть постраничные растровые файлы с разными настройками качества под разные цели (максимальное качество для хранения в фондах и OCR, компрессированные файлы для веб портала и т.д.) Обработка должна быть максимально гибкой и настраиваться под каждый проект индивидуально.

В архитектурном смысле продукт должен представлять собой распределенную модульную систему.

Система должна легко масштабироваться как для индивидуального использования, например, в небольших учреждениях, где для осуществления сканирования и обработки достаточно задействовать одного сотрудника, так и в крупных учреждениях, где сканированием, обработкой и контролем качества могут заниматься десятки сотрудников.

Система должна быть независимой от конкретных моделей сканеров и обеспечивать возможность параллельной обработки большого количества проектов одновременно.

Система должна обеспечивать надежный контроль качества автоматической обработки и возможность «отката» обработки отдельных изображений на любом шаге до любого из предыдущих.

Насколько нам известно, библиотека рассматривала несколько программных продуктов зарубежных производителей для решения стоящих задач, но набор функций, лицензионная политика их производителей и конечная стоимость решений привели к исключению данных вариантов из рассмотрения.

Наша компания на этот момент обладала программным инструментарием, который позволял решить бОльшую часть обозначенных задач. Однако, продукта, который бы объединял все инструменты в одном продукте и вписывался в рабочие процессы, инфраструктуру библиотеки и обеспечивал бы необходимую производительность, у нас на тот момент не было. Кроме того, часть инструментов обработки, созданные в ALANIS Software, потребовали значительных доработок для получения наиболее качественных результатов на предоставленных нам сэмплах.

У нас было примерно полгода для создания продукта с использованием наших разработок и представлением его на рассмотрение РГБ. Мы уложились в сроки и первая версия продукта Book Scan Processing была подготовлена к концу 2015 года.

С этого момента система постоянно совершенствуется на основе обратной связи, которую мы получаем от Российской Государственной Библиотеки.

Решения

Функции обработки

Предлагаю более детально рассмотреть ключевые особенности системы ALANIS Book Scan Processing (BSP).

Для этого, давайте сначала ознакомимся с примерами сканов, которые система должна обрабатывать. Сканы, которые были нам предоставлены, были получены с помощью книжных сканеров различных моделей и очень отличались друг от друга, но при этом все они имели общие характерные черты:

1. Каждый скан являлся половиной разворота книги
2. Помимо отсканированной страницы скан содержал линию корешка и часть соседней страницы, а также области за пределами страницы (стол сканера, внутренние края обложки, обрез книги)
3. Многие сканы имели неравномерное освещение. Более темные области у корешка книги и с краев страниц из-за выпуклости страниц при сканировании. Особенно ярко это проявляется у книг с большим количеством страниц.
4. Следствием предыдущего пункта являются геометрические искажения содержимого, текстовые строки превращаются в кривые, особенно явно это отмечается рядом с корешком книги.
5. Многие сканы также имели загрязнения, артефакты, появившиеся из-за возраста и качества бумаги

Инструменты автоматической обработки

Чтобы на выходе получить идеально выровненные, правильно обрезанные страницы наилучшего визуального качества программисты ALANIS Software включили в систему BSP набор фильтров автоматической обработки (я приведу их в обычном порядке применения, хотя в зависимости от характеристик сканов конкретного источника порядок фильтров может варьироваться):

- Устранение перекоса

Данный фильтр имеет два режима:

- выравнивание по тексту
- выравнивание по краям страниц (используется для документов с большим количеством графической информации и относительно небольшим количеством текста, например, для художественных альбомов)

– Обрезка по значимой информации

Данный фильтр был специально разработан нами с нуля для Book Scan Processing и потребовал значительных временных затрат на тестирование и доводку. Он представляет собой так называемую «плавающую» обрезку с фиксированным размером рамки. Это означает, что программа анализирует каждый скан, находит на нем область значимой информации (текст, иллюстрации, пометки, которые необходимо сохранить на обрезанном скане), описывает данную область прямоугольником. Центр данной области будет являться и центром прямоугольника обрезки. Размер прямоугольника обрезки задается пользователем для всех страниц проекта. Таким образом, в результате обрезки мы получаем набор страниц одного размера с одинаковым положением области значимой информации. В процессе работы над фильтром мы ввели ещё один параметр помимо размера рамки обрезки, настраиваемый пользователем. Для повышения точности работы фильтра мы ввели понятие максимально допустимого размера области значимой информации, что позволило избежать ситуаций, когда текст с соседней страницы мог определяться как принадлежащий к области значимой информации. Для наглядности и удобства пользователя данный параметр выглядит как «резиновая» пунктирная рамка внутри рамки обрезки поверх исходного скана.

Я хотел бы отметить, что обрезка сканов в существующем рабочем процессе РГБ занимала огромное количество времени и выполнялась вручную с помощью графического редактора Photoshop. Оператор открывал каждый скан и вручную приблизительно позиционировал рамку так, чтобы все страницы выглядели одинаково после обрезки. Внедрение автоматического фильтра «умной» обрезки кардинально сократило время на обработку одной страницы.

– Выравнивание строк

Фильтр определяет наличие текстовых строк на скане без использования OCR и трансформирует скан таким образом, чтобы все строки были строго горизонтальны. Особое внимание при разработке фильтра было уделено обработке областей прилежащих к линии корешка книги, так как здесь, как правило, встречаются наиболее сильные искажения.

Стоит отдельно отметить, что помимо полностью автоматического фильтра BSP предоставляет возможность ручной коррекции кривизны строк на этапе контроля качества. Инструмент представляет собой сетку поверх скана, шаг ячеек сетки устанавливается пользователем в зависимости от плотности текста и количества «волн» на конкретном скане. Чтобы внести коррективы достаточно потянуть за определенные узлы сетки в нужном направлении, изображение «потянется» за сеткой.

– Выравнивание освещенности

Планетарные сканеры требуют особого внимания к освещению сканируемого источника, особенно, если сканируется ветхие или особо ценные материалы, не допускающие использования прижимного стекла. Далеко не всегда удается достичь равномерной засветки всей площади страниц при сканировании, в частности, из-за изгибов страниц. Минимизировать огрехи освещения позволяет разработанный нашей компанией фильтр, который уравнивает уровень засветки разных частей страницы, приводя их к среднему значению. Обрабатываются как пересвеченные, так и недосвеченные части скана.

– Устранение фактурности бумаги

Данный фильтр оказывает в большей степени «косметическое» действие и позволяет автоматически «отретушировать» фон скана, выражаясь «доцифровым» фото языком. Фильтр оказывает эффект размытия, при этом текстовая информация сохраняет свою четкость. Кроме того, фильтр может служить промежуточным звеном улучшения изображения перед бинаризацией (перевод в ч/б), его применение уменьшает количество мелкого «мусора» на черно-белых изображениях.

– Выравнивание фона по шаблону

Использование данного фильтра позволяет достичь однородного цвета фона на всех страницах обрабатываемого источника, что очень актуально, например, при создании электронных версий книг. Цвет фона может значительно отличаться как из-за разного качества бумаги внутри одного источника (при длительном хранении эффект проявляется особенно явно), так и в результате сканирования. Например, нередки ситуации, когда освещение левой и правой страницы при бесконтактном сканировании имеет отличия. Также нередки ситуации, когда сканирование одного источника происходит в несколько этапов или на разных сканерах, что также непосредственно влияет на цветовой оттенок и яркость фона сканов.

Для сохранения присущего бумаге конкретной книги цветового оттенка можно выбрать одну из страниц в качестве шаблонной. Тогда средние цветовые характеристики этой страницы будут использованы для унификации цвета всех страниц пакета. Выбранный цвет можно скопировать в виде HEX кода для использования на других книгах, например, если сканируется собрание сочинений одного автора.

– Улучшение иллюстраций

Уникальность фильтра состоит в том, что он позволяет автоматически «вытянуть» даже очень слабо читаемые изображения. Подобные алгоритмы используются, например, для оптимизации качества снимков аэрофотосъемки, сделанных в условиях недостаточной видимости, например, в сумерках или в условиях тумана. Фильтр весьма актуален при создании электронных версий старинных альбомов по искусству и богато иллюстрированных изданий.

Инструменты ручной обработки

Какой бы ни была автоматическая обработка, она не может быть абсолютно безупречной на бесконечном разнообразии сканируемых источников, поэтому системе необходимы удобные инструменты ручного контроля обработки и коррекции.

Важно отметить, что ручная обработка не всегда означает коррекцию каждого отдельного изображения в полностью ручном режиме. Цель этапа контроля качества – исправление недочетов автоматической обработки с минимальными временными затратами. Поэтому на данном этапе оператору программы доступны как ручные инструменты, так и весь набор ранее рассмотренных автоматических фильтров.

Для ситуаций, когда автоматическая обработка в принципе не дает нужного результата из-за особенностей изображения, предусмотрены ручные инструменты:

– Ручная обрезка страницы

Это стандартный инструмент обрезки рамкой фиксированного размера. Особенность его применения в том, что он заменяет собой фильтр автоматической обрезки с сохранением заданных размеров при загрузке изображений в модуль контроля качества. При этом, его также можно применять отдельно.

– Ручное выравнивание строк

Инструмент представляет собой сетку поверх изображения, позволяющую «по месту» корректировать искривления текстовых строк, перемещая узлы сетки. Изображение «плывет» за соответствующим узлом. Количество (плотность) ячеек сети устанавливается пользователем, что позволяет точечно корректировать самые разнообразные искажения.

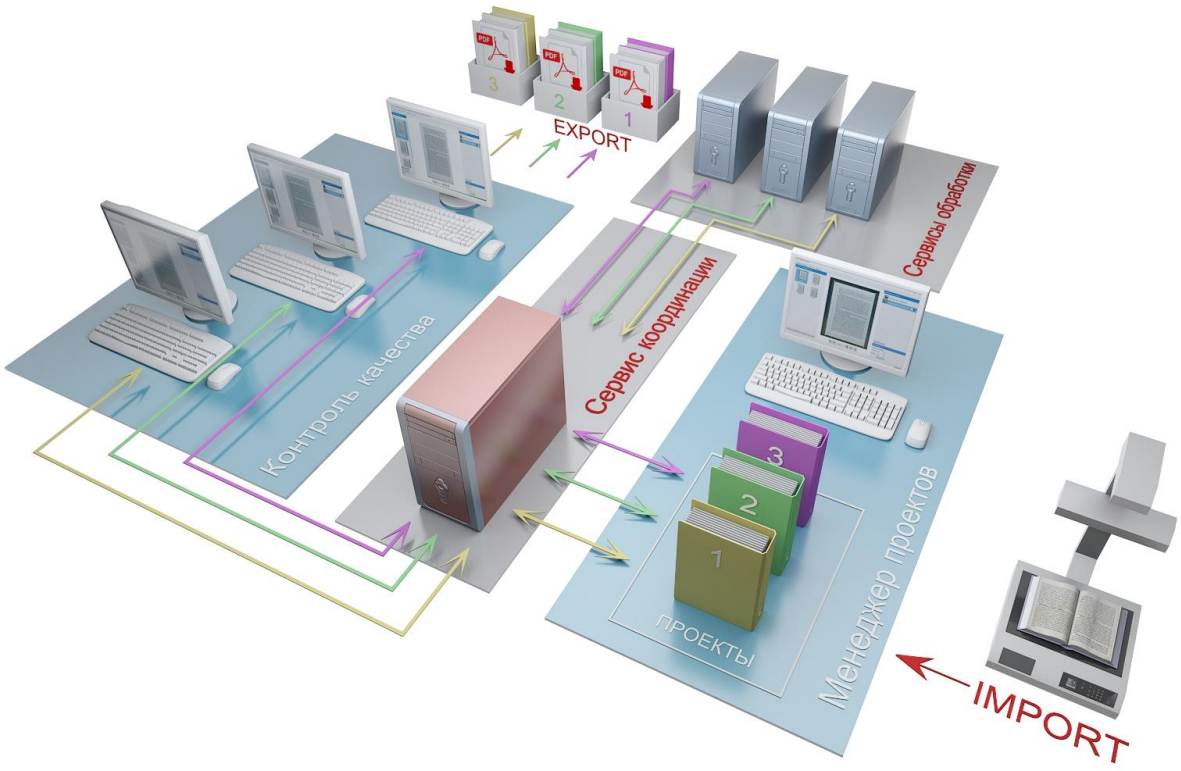
– Поворот на заданный угол

Максимально наглядный и интуитивный инструмент позволяет пользователю выбрать произвольную линию на изображении (это может быть строка текста, обрез иллюстрации или край страницы) которая должна быть «линией горизонта» на обработанном изображении, провести с помощью мыши параллельную линию. Как только кнопка мыши будет отпущена, изображение повернется соответственно.

Архитектура системы

Book Scan Processing состоит из нескольких сервис компонентов, ответственных за распределение задач, осуществление обработки и экспорта, сбор статистики и коммуникацию между всеми компонентами системы. Пользовательские компоненты позволяют настроить и запустить процессы обработки, а также проконтролировать качество обработки. Все компоненты системы работают в одной локальной сети.

В общем виде архитектура системы представлена на следующей схеме:



Определение проекта

Единицей обработки в системе BSP является проект. Проект представляет собой набор сканов одного печатного источника, например, книги или журнала. Все изображения одного проекта должны иметь, (и, как правило, имеют) схожие характеристики: размер страницы, условия сканирования. Это позволяет применить ко всем изображениям проекта один набор функций обработки. Помимо этого, настройки проекта включают в себя имя проекта, индивидуальные сетевые пути к папками импорта исходных изображений, экспорта результатов обработки, а также настройки форматов сохранения результатов обработки. Отдельно стоит отметить, что каждый проект обязательно создается от имени конкретного пользователя системы и назначается конкретному пользователю, который осуществляет контроль качества обработки. Это позволяет адресно распределять задачи обработки и отслеживать производительность работников. При необходимости, возможно переназначение проектов от одного пользователя к другому.

Жизненный цикл проекта детально документируется в хронологическом порядке с использованием событийной модели. Документируются события разных типов: этапы обработки, ошибки обработки, как на уровне проекта, так и на уровне обработки отдельных изображений и даже фильтров обработки.

Серверные компоненты BSP

– Сервис координации

В центре всей системы находится серверный компонент сервис координации, можно сказать, что это «мозг» для нервной системы BSP. Все остальные компоненты системы общаются

только с ним. Сервис координации знает всё обо всех модулях системы и обеспечивает распределение и прохождение задач импорта изображений, обработки и экспорта в системе. Сервис координации может быть установлен на любом компьютере в сети, как изолированно, так и вместе с другими компонентами системы.

Наличие одного сервиса координации определяет один рабочий экземпляр Book Scan Processing. При этом, в одной локальной сети может существовать несколько экземпляров Book Scan Processing, каждый со своим сервисом координации. Это удобно при установке в крупных учреждениях, особенно, если разные отделы осуществляют разные задачи оцифровки.

Ещё более удобно то, что отдельные модули BSP (сервисы обработки, модули управления проектами и контроля качества) можно «перебрасывать» от одного экземпляра BSP в другие при необходимости, например, для наращивания мощностей по обработке. Никаких глубоких знаний по администрированию системы при этом не требуется.

– Сервис обработки

Данный компонент системы содержит все необходимые фильтры и непосредственно осуществляет все операции по пакетной обработке изображений согласно настройкам поступающих проектов.

Один экземпляр системы BSP может иметь несколько сервисов обработки, расположенных на разных компьютерах, их количество определяется потребностями пользователей и условиями конкурентной лицензии продукта. Сервис обработки может располагаться и на одной машине с другими компонентами системы, однако рекомендуется выделять отдельные производительные машины, особенно для обработки больших объемов файлов за короткое время. Система автоматически загружает все доступные сервисы обработки, таким образом, балансируя нагрузку.

Десктоп (пользовательские) компоненты BSP

– Модуль настройки проектов

Данный модуль предназначен для создания, настройки, запуска и отслеживания жизненного цикла проектов обработки.

Главное окно программы позволяет отслеживать состояние существующих в системе проектов, а также создавать, удалять проекты и останавливать обработку отдельных проектов.

Помимо собственно управления проектами пользователь с правами администратора может заводить в системе новых пользователей, редактировать их данные, удалять их, а также назначать других пользователей системы на проекты в качестве операторов контроля качества. «Обычный» пользователь также может создавать проекты, однако оператором контроля качества он может назначить только самого себя, что удобно при автономном режиме работы. Кроме того, «обычный» пользователь не видит проекты, назначенные другим пользователям, и у него нет привилегий создания и редактирования данных других пользователей системы.

Модуль настройки проектов может быть установлен в любом количестве экземпляров в сети и может подключаться к разным экземплярам BSP, если их несколько.

Пользовательский интерфейс окна создания/настройки проекта организован согласно логике жизненного цикла проекта и состоит из следующих вкладок:

- Общие свойства

  Здесь задаются имя проекта и пользователь системы, ответственный за контроль качества.
- Импорт

  Указывается одна или несколько папок, из которых будут импортированы изображения для обработки. При этом папки могут содержать вложенные подпапки. Путь к папкам задается в сетевом формате (UNC), если локальная папка не является доступной по сети при указании её как папки импорта, для неё автоматически будет создана сетевая папка с общим доступом. Таким образом, сервисные компоненты BSP, которые обычно располагаются на специально выделенных компьютерах в сети, могут напрямую получать изображения для обработки.

- Обработка
  Данная вкладка содержит гибкий визуальный редактор настройки автоматических фильтров обработки на сэмпловых изображениях проекта. Здесь пользователь выбирает несколько файлов из начала, середины и конца отсканированной книги и настраивает набор фильтров проекта таким образом, чтобы получить качественный результат обработки на всех страницах.
- Экспорт
  В данной вкладке пользователь указывает папку сохранения результатов обработки и конфигурирует параметры качества двух форматов экспорта: PDF и одного из растровых форматов (JPG ; JPG2000, TIFF, PNG).
  В зависимости от целей оцифровки характеристики файлов могут различными. Файлы с высокой компрессией подходят для публикации в электронных коллекциях, файлы в максимальном качестве – для целей сохранения в библиотечных фондах в качестве цифровых мастер копий.

– Модуль контроля качества

Данный модуль всегда запускается от определенного пользователя системы BSP. Таким образом, пользователь автоматически получает с сервиса обработки проект, в котором он назначен оператором контроля качества. Если для авторизовавшего пользователя нет активных проектов, модуль будет находиться в режиме ожидания. В таком случае, пользователь всегда может создать проект обработки самостоятельно с помощью модуля управления проектами и назначить себя оператором контроля качества.

Интерфейс модуля контроля качества по большей части идентичен вкладке «Обработка» модуля управления проектами.

Изображения поступают в модуль по мере обработки на сервере, а, значит, оператору не нужно дожидаться момента окончания обработки всех изображений проекта, что позволяет избежать простоев в работе.

Поступающие изображения имеют полную историю обработки в виде списка примененных фильтров (исходное изображение и все промежуточные состояния обработки доступны оператору), где каждый фильтр можно редактировать. Таким образом, при обнаружении проблемы достаточно точечного действия по исправлению конкретного примененного фильтра обработки, и не нужно переделывать всю работу с нуля.

Кроме того, если в процессе обработки конкретного изображения какой либо из фильтров выдает ошибку (например, угол поворота страницы в фильтре устранения перекоса превысил допустимое значение или фильтр исправления кривизны строк не смог определить достаточное количество строк для вычисления необходимой матрицы коррекции), ошибка будет видна оператору в виде предупреждающего значка поверх миниатюры изображения в галерее, что также обеспечивает адресность операций коррекции.

Модуль также позволяет мгновенно откатиться к состоянию изображения по результатам автоматической обработки, если действия оператора не привели к положительному эффекту.

Дополнительное средство оптимизации скорости работы оператора контроля качества – повторная обработка, в ходе которой выборка изображений с новыми настройками отсылается обратно на сервер. В то время, пока проходит повторная обработка на сервере, оператор продолжает проверку других страниц пакета. Обработанные повторно изображения поступают в модуль контроля качества по мере готовности для контроля, либо отправляются непосредственно на фазу экспорта, если по мнению оператора повторная обработка не требует дополнительной проверки.

Выводы и перспективы

Система Book Scan Processing на данном этапе своего короткого существования уже позволяет значительно сократить долю ручного труда при подготовке сканов в проектах оцифровки библиотечных фондов. В процессе сотрудничества с Российского Государственной Библиотекой компания «АЛАНИС Софтвер» приобрела ценнейший опыт практической реализации легко масштабируемого и тиражируемого промышленного продукта для той сферы деятельности профессионального библиотечного сообщества, которая незаслуженно обходится вниманием отечественных ИТ

компаний. Практически все библиотеки, с которыми мы общались, пользуются для оптимизации изображений программой Adobe Photoshop, которая предоставляет огромные возможности для работы с растровыми изображениями, но мало подходит для целей библиотечной оцифровки. Использование Photoshop в массовой подготовке сканов можно сравнить с забиванием гвоздей микроскопом. «АЛАНИС Софтвер» же предлагает решение, построенное на основе требований тех, кто ежедневно обрабатывает сотни и тысячи сканов. Поэтому, в Book Scan Processing есть все необходимое для производительной и качественной обработки. Мы постарались учесть максимум пожеланий пользователей системы в части удобства работы с модулями, поэтому интерфейсы модулей имеют минималистичный «спокойный» дизайн и требуют минимум «телодвижений» для выполнения частых действий. Мы сознательно отказались от создания интерфейсов в стиле «кабина пилота авиалайнера» и минимизировали количество цветов и декоративных элементов в окнах программы, поскольку мы знаем, сколько часов в день проводит оператор, глядя в монитор.

Отдельно хотелось бы отметить возможности масштабирования системы под нужды конкретных учреждений, отделов и отдельных проектов оцифровки. Переформатирование системы под новую конфигурацию по сути заключается в запуске модулей системы на нужных компьютерах и указании в запущенных модулях IP адреса сервиса координации, с которым модули должны работать.

Закончено ли развитие Book Scan Processing? Конечно, нет. Мы рассматриваем текущее состояние системы как базовую конфигурацию, которая будет дополняться новыми функциональными модулями, существующие модули будут совершенствоваться.

В первую очередь, я хотел бы сообщить, что новая версия Book Scqn Processing, учитывающая накопленный нами опыт за время опытной эксплуатации в Российской Государственной Библиотеке, выходит в июле этого года. Мы кардинально переработали архитектуру системы, оптимизировали взаимодействие компонентов системы и улучшили пользовательское качество продукта, которое емко описывается английским термином “user experience”.

В каком направлении будет развиваться система? В первую очередь, мы расширим набор функций обработки, доступных в BSP. АЛАНИС Софтвер обладает собственным арсеналом из более чем 40 функций обработки изображений, только треть из них на данный момент включены в состав BSP. Архитектура системы позволяет легко интегрировать новые функции.

Поскольку компания работает как на российском, так и на международном рынке, в дальнейшем BSP будет иметь англоязычный интерфейс.

Мы планируем расширить функциональность системы модулем ввода метаданных с поддержкой стандартов библиографических данных.

Следующим логичным шагом в развитии системы является включение в состав BSP сервиса распознавания текста и расширения форматов экспорта такими востребованными текстовыми форматами как HTML, ALTO XML, и т.д. Учитывая тот факт, что BSP изначально создан для нужд академического сообщества, мы считаем многообещающим использование системы Tesseract в качестве системы распознавания. Tesseract широко используется в международных проектах сохранения исторического наследия и поддерживает большое количество современных и старых языков, благодаря исследователям из разных стран, создающим языковые пакеты, например, для немецкой готики, древнегреческого, староанглийского и старофранцузского языков, и т.д.

Если система BSP предоставит удобный способ подключения языковых пакетов Tesseract, это позволит максимально адаптировать систему под нужды исследовательских проектов, проектов сохранения культурно-исторического наследия.

Подытоживая сказанное, мы надеемся, что наш продукт будет полезен в каждодневной работе библиотек по всей России и за рубежом. Мы, в свою, очередь, готовы совершенствовать систему, чтобы Ваш труд по созданию цифровых ресурсов был максимально производительным и легким.