УДК 004;004.93

В.В. СТАРОВОЙТОВ, Объединенный институт проблем информатики НАН Беларуси

О ЦИФРОВОЙ РЕСТАВРАЦИИ ИСТОРИЧЕСКИХ ТЕКСТОВЫХ ДОКУМЕНТОВ

Рассматриваются основные проблемы реставрации старых текстовых документов и подходы к их решению при реставрации изображений этих документов методами информационных технологий. Документы-первоисточники при этом не подвергаются изменениям, а их цифровые копии можно модифицировать с ориентацией на различные применения и согласно разным уровням обработки.

The main problems of old text document restoration and approaches to their solution during restoration of images of these documents by methods of information technology are considered. Primary source documents are not changed, but their digital copies can be modified with orientation on different applications and according to the different levels of processing.

Введение

Традиционная реставрация старых культурных ценностей процесс очень кропотливый и медленный, например, фреску «Тайная вечеря» Леонардо да Винчи реставрировали 21 год. Главный принцип такой реставрации – не навредить оригиналу. Часто к ней прибегают, когда разрушение картины, памятника или другого объекта культурного наследия находится под угрозой утраты целостности и не позволяет его использовать и демонстрировать.

В последние годы многие национальные библиотеки и архивы переводят свой фонд в цифровое представление. Многие исторические документы (книги, чертежи, карты, рукописи и т.п.) в силу ветхости или уникальности доступны только ограниченному кругу специалистов. Их перевод в цифровое представление обеспечивает возможность доступа к ним широкому кругу читателей, причем даже дистанционно. Однако многие старые документы имеют искажения и повреждения, что затрудняет их чтение и изучение.

В проекте ГОСТ РФ «Электронные документы. Основные виды, выходные сведения, технологические характеристики» детально описывается понятие электронного документа (ЭД), как созданного программными средствами целостного электронного объекта, предназначенного для непосредственного восприятия человеком. Выделяются три вида электронных документов:

- текстовый ЭД, содержание которого составляет читаемая информация преимущественно в виде слов;
- графический ЭД, содержание которого составляет визуальное представление объектов;
- звуковой (аудио) ЭД.

По представлению данных выделяются:

- оригинальный ЭД (созданный впервые),
- копия ЭД,
- цифровая копия аналогового документа,
- трансформированный ЭД (переведённый из одной знаковой системы в другую, например, с помощью распознавания текста, синтеза речи).

В данной работе будем ориентироваться на создание графического ЭД являющегося отреставрированной цифровой копией реального документа. Основным отличием создания цифрового представления исторического документа является отсутствие цифрового оригинала, т. е. оригинальный документ создавался без участия вычислительной техники.

Электронные документы в цифровых фондах библиотек

Основные организации хранящие исторические документы – это архивы и библиотеки. Следует отметить, что во многих странах библиотеки активно формируют свой цифровой фонд. Рассмотрим основные направления цифрования текстовых документов по регионам.

В мире. В одной из самых больших библиотек мира – Библиотеке конгресса США – по данным на 2007 год было оцифровано 10% из 142 млн книг и документов. В первую очередь в электронный вид были переведены документы, датированные до 1923 года, поскольку на них отсутствуют авторские права. Ежедневно в ней сканируется от 75 до 200 документов. По подсчетам специалистов, для оцифровки всего ее фонда (142 млн. объектов) потребуется не одно десятилетие [1]. В Национальной библиотеке Франции с 1997 года реализуется проект Gallica (gallica.bnf.fr), в рамках которого отсканированы, переведены в цифровой формат и выложены в Интернет порядка 80 тысяч книг и 70 тысяч изображений. Британская национальная библиотека в 2005 году заявила о начале реализации проекта по переводу фонда в электронный вид. К 2020 году около 90% научной продукции британских ученых переведут в цифровой формат, и они будут доступны через Интернет. В марте 2005 года начала свою работу Европейская цифровая библиотека (The European Library). Европейская библиотека – это совместный некоммерческий проект 48 национальных библиотек Европы. Россию в этом проекте представляют Российская государственная и Российская национальная библиотеки. Сегодня Европейская цифровая библиотека предлагает доступ к 150 млн. документов на 35 языках, хранящихся в 48 национальных библиотеках европейских стран [2]. К сожалению Беларусь не присоединилась к этому проекту.

Наряду с национальными и континентальными проектами электронных библиотек в 2005 г. ООН учредила международный проект «Мировая цифровая библиотека» (World Digital Library). Основная идея заключалась в создании на базе сети Интернет легкодоступной коллекции сокровищ Мировой культуры, способствуя укреплению межкультурных связей и взаимопониманию. «Мировая цифровая библиотека» официально заработала в апреле 2009 года. Содержание веб-сайта, включает библиотеки и архивы со всего мира, доступно на семи языках – арабском, китайском, английском, французском, португальском, русском и испанском.

В России. Ряд библиотек России также создает и хранит электронные копии различных документов. Российский проект «Национальная электронная библиотека» (НЭБ) разрабатывается ведущими российскими библиотеками при поддержке Министерства культуры Российской Федерации с 2004 г. На сегодняшний день в проекте участвуют около 60 библиотек-партнеров, среди которых: Российская государственная библиотека (РГБ), Российская национальная библиотека (РНБ), Государственная публичная научно-техническая библиотека (ГПНТБ России). Общий объем электронных документов НЭБ сегодня составляет 15 миллионов страниц в электронном виде, и библиотека постоянно пополняется. Одно из направлений – старопечатные книги включает цифровые копии уникальных изданий, вышедших до 1830 г. Все документы хранятся в формате pdf.

17 декабря 2008 года Российская государственная библиотека объявила о завершении проекта по созданию электронного хранилища книг, созданного в рамках реализации концепции Национальной электронной библиотеки. Хранилище объединило редкие книги и рукописи, периодические издания и ноты, собранные со всех основных российских библиотек. В него вошли также книги Президентской библиотеки и электронной библиотеки диссертаций [3].

Основу электронной библиотеки «Земля Владимирская» составляют электронные копии редких и краеведческих изданиях из фондов библиотек региона. В электронной коллекции содержится «Банк правовых актов Владимирской губернии» для в цифровом формате. Этот ресурс, наиболее полно представляющий историю становления и развития государственности во Владимирской области со времени основания губернии в 1778 г. до конца 1990-го г. Проект реализуется с 2009 г. [4].

В Беларуси основными организациями-хранителями исторических документов являются:

- Национальный исторический архив Беларуси (НИАБ),
- Национальная библиотека Беларуси,
- Белорусский научно-исследовательского центр электронной документации.

Существует «Белорусская цифровая библиотека» (БЦБ) – частная некоммерческая интернет-библиотека Республики Беларусь. Коллекция библиотеки содержит более 100 000 публикаций и пополняется в автоматическом режиме пользователями сайта: каждый посетитель может разместить свои научные и литературные труды на сайте (http://library.by/).

Следует отметить что большинство ЭД хранящихся в белорусских библиотеках – это современные текстовые документы, подготовленные с помощью компьютеров и преобразованные (в основном) в формат *pdf* из других тестовых форматов с помощью графических редакторов.

Внимание архивистов в нашей стране больше сосредоточено на вопросах цифрового копирования традиционных документов, уже хранящихся в архивах. Если копирование книжной продукции направлено прежде всего на обеспечение доступа (и потому актуальна проблема авторского права), то для архивов, а также рукописных отделов библиотек первостепенную роль играет возможность обеспечить лучшую сохранность оригиналов, обычно существующих в единственном экземпляре [5]. К сожалению, работы по оцифровыванию, а тем более по цифровой реставрации исторических текстовых документов в нашей стране не выполняются.

Реставрация старых текстовых документов

Старые текстовые документы не имеют исходных цифровых форматов и чаще всего представлены оригинальными бумажными документами. По способу нанесения текста бумажные документы можно разделить на три основные группы: рукописные, машинописные, печатные. [6]. Рукопись почти полтора тысячелетия являлась единственным способом исполнения бумажных документов (II–XVI вв. н.э.). В этот период документы не только исполняли, но и копировали рукописным способом. Несколько позднее применялось факсимильное ксилографическое копирование рукописных книг (VII–VIII вв. – Китай и Япония; XV в. – Европа). В XV в. появляется буквонаборное устройство, печатный станок и книгопечатание (1440 г.). В XVI в. рукописные книги вытесняются печатными, однако рукопись, как способ исполнения официальных документов, сохраняет свое значение до конца XIX в. В конце XIX в. появляется машинописный способ исполнения документов, играющий основную роль до конца XX века. В настоящее время основной способ создания документов – компьютерный.

Состав средств нанесения символов постоянно изменялся. До 70-х гг. XIX в. для написания текста использовали только природные вещества: неорганические пигменты, органические красители растительного и животного происхождения. С конца XIX в. природные красители были заменены синтетическими. С момента появления машинописи (1887 г.) ее технология принципиально не изменилась. Машинописный текст переносился на бумажный носитель с пропитанной краской ленты (1-й экз.) или с копировальной бумаги (2–5-й экз.). Краска для лент содержала в два раза больше красителей и жидких масел, чем копировальная. Поэтому цветовая насыщенность, глубина прокрашивания бумаги, устойчивость к истиранию у текста первого экземпляра значительно выше, чем у копий.

Машинописные тексты водостойки, черные рецептуры светостойки. Цветные машинописные тексты выцветают на свету, нестойки к химическим веществам (щелочам, окислителям), расплываются в неполярных органических растворителях (толуоле, бензоле и т. п.). Краски для печати состоят в зависимости от назначения продукции из сажи, олифы, минеральных масел, каменноугольных смол. Краски глубоко проникают в бумагу, печатный текст свето- и водостоек. При действии органических растворителей может появляться ореол от вымывания смол или красителей.

В современных электрофотографических аппаратах текст наносится на бумагу с помо-

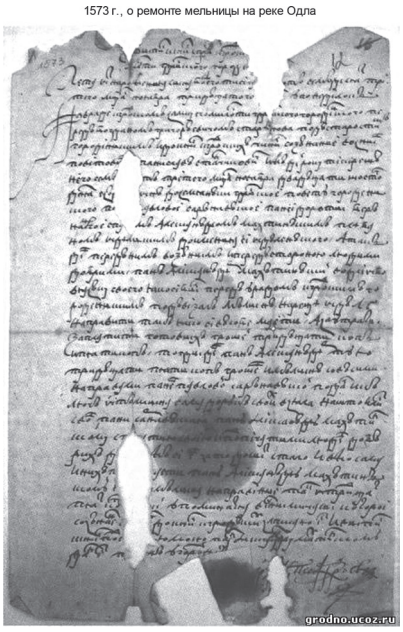


Рис. 1. Цифровое изображение белорусского документа 1730 года: неравномерный фон, отсутствуют фрагменты, имеются пятна, потертости, нечеткие символы, более поздние пометки

щью специального порошка (тонера) и закрепляется термическим способом. Черный тонер состоит из сажи и термопластичного полимера (идитол, полистирол и т. п.). Электрографический текст водостоек, светостоек, имеет такую же прочность к истиранию, как тушевой или машинописный текст (1-й экз.). Текст термопластичен, повреждается при нагревании, расплывается в органических растворителях.

С течением времени и вследствие не всегда удовлетворительного хранения носители таких документов подвержены старению и повреждению, а текст и графика теряют свое визуальное качество. Со временем старые документы становятся все менее доступны читателям. В Беларуси существует государственное учреждение «Центральная лаборатория микрофильмирования и реставрации документов Национального архивного фонда Республики Беларусь» (ЦЛМРД НАФ РБ). Оно занимается традиционной реставрацией документов, детали которой подробно описаны в [6].

Традиционная реставрация проводится с целью физического восстановления документов, разрушенных при старении [7]. Ее основные задачи:

- восстановление физической целостности документа,
- восстановление прочности бумажного носителя,
- устранение причин и последствий старения документа,
- устранение причин, вызывающих искаженное эстетическое восприятие документа.

Реставрационное вмешательство всегда сопряжено с опасностью повредить документ.

Виды повреждений исторических документов

Механические повреждения, для которых типичны четкие линии повреждений и отсутствие протяженных переходных зон от целого к разрушенному (обрывы, разрывы, проколы, порезы, места сгибов и т. п.). Механические повреждения не приводят к изменению химического состава и свойств объекта и устраняются способами физической реставрации (соединение разрывов, восполнение недостающих частей, долив бумажной массы и т. п.).

Повреждения насекомыми по характеру действия являются механическими. Типичным является сочетание точечных, линейных, кружевных отверстий с измельченными в труху отдельными, чаще всего краевыми местам листов. Насекомые не выделяют химических веществ в местах повреждений и поэтому устранение этих дефектов проводится так же, как механических.

Повреждения плесневыми грибами имеют характерные внешние признаки: бумага по всему листу или крупными зонами разрушена, стала ломкой и хрупкой, побурела; зоны поражения имеют пигментные пятна различных, чаще всего оранжево-желтых и коричневых цветов; после сильного плесневения видны налеты мицелия и порошка спор; бумага в местах поражения плохо смачивается водой, кислотность ее повышена.

Повреждения химические могут быть двух видов: общие и локальные. При общем химическом поражении потеря прочности, ветхость, желтизна бумаги, выцветание текста равномерны и примерно одинаковы по всей площади листа. Такое повреждение является обычно следствием длительного темнового старения, кратковременного действия тепла или света.

Для локального поражения характерны повреждения в зонах или отдельных местах. Повреждения вызываются кислотами, щелочами, солями, попавшими на бумагу случайно или вместе с наклейками, чернилами. Действие кислых газов заметнее на краях листа, имеющих повышенную хрупкость, желтизну, кислотность.

В некоторых случаях сильные химические повреждения напоминают по внешнему виду плесневые, но не имеют типичных биологических признаков – пигментации, налетов спор и мицелия.

Повреждения водой определяются по следам подмочки, размытому тексту, деформации бумаги. Часто в местах подмочки заметны следы плесневения, особенно в корешках дел. Намокание документов нередко сопровождается попаданием на бумагу грязи, различных солей.

Повреждения огнем имеют типичные внешние признаки: следы обугливания, сажевые загрязнения, побуревшая, хрупкая, рассыпающаяся бумага, обесцвеченный, поврежденный текст.

Действие разрушающих факторов – света, тепла, воды, химических веществ, плесневых грибов изменяет структуру бумаги, приводит к деструкции органических веществ документа

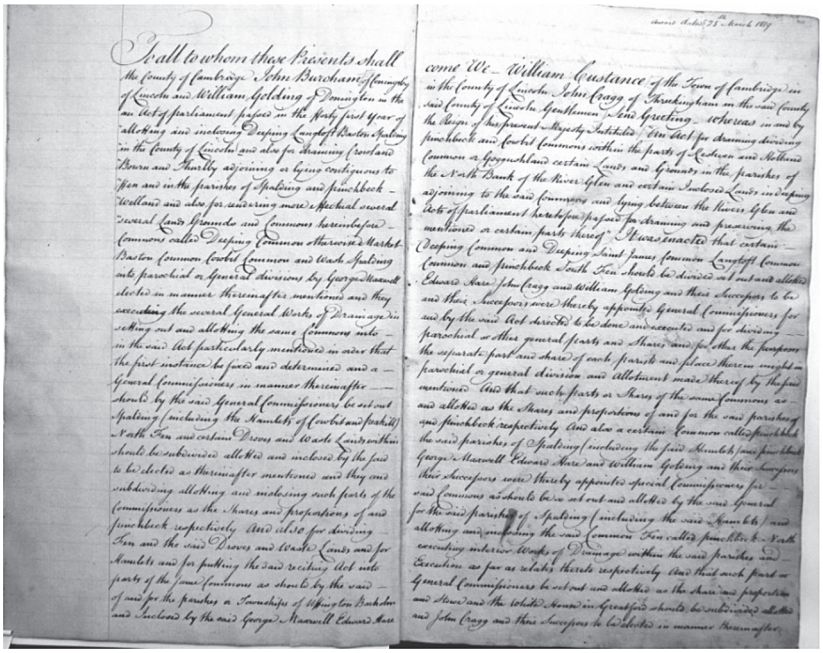


Рис. 2. Геометрическое искажение формы строк, неоднородный фон возникшие в результате оцифровывания документа, а также пятна на странице

и образованию продуктов распада. Резко уменьшается прочность бумаги, ее сопротивляемость любым воздействиям. Процессы старения в ослабленной бумаге протекают ускоренно, документы легче поражаются биологическими вредителями, не выдерживает обычных нагрузок. Выцветает и легче повреждается текст.

Можно выделить следующие дефекты бумажных текстовых архивных документов [6, 8]:

- прозрачные и непрозрачные пятна на страницах,
- загрязнение поверхности,
- царапины и надрывы страниц,
- перегибы бумаги (или другого материала-основы),
- дыры и оборванные края и углы документа,
- деформация документа (кроме перегибов),
- выцветание материала-основы, пожелтение бумаги,
- неравномерность фона,
- плесень,
- выцветание текста,
- рукописные пометки на страницах,
- штампы, печати,
- надписи, проступание с обратной стороны листа.

Традиционная реставрация проводится с целью физического восстановления или сохранения основы документов, разрушенной или измененной с течением времени.

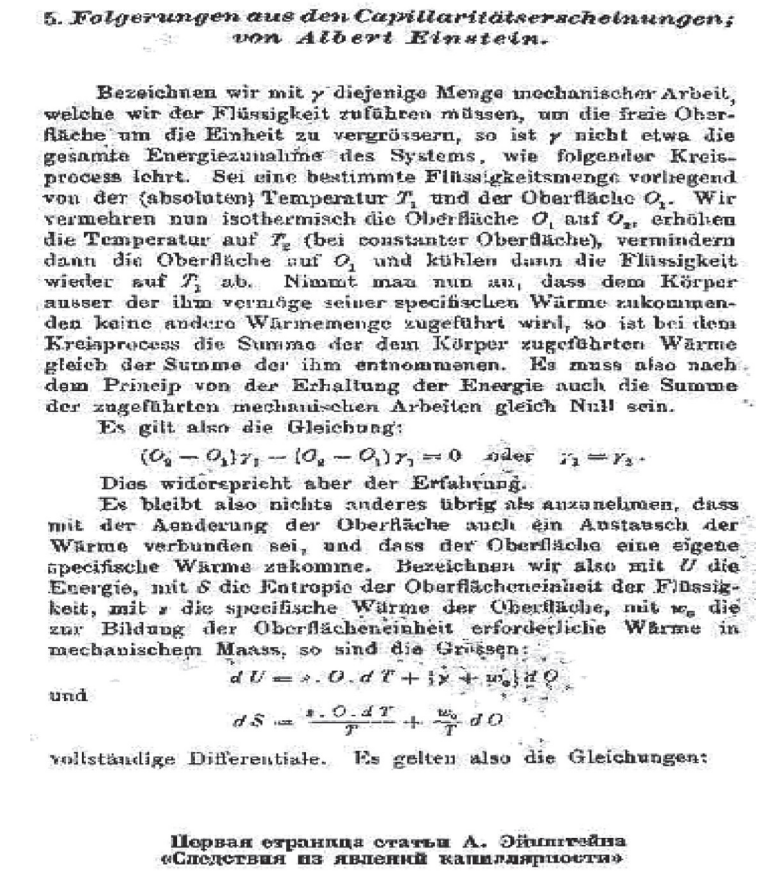
Цифровая реставрация старых документов

Создание цифровых копий исторических документов и их цифровая реставрация позволит сделать эти документы доступными для широкого круга людей, а более качественное представление информации, которую они несут, сохранит их для будущих поколений. Некоторые методические и технические требования к оцифровке исторических документов описаны в работах [9–10], но задачи и алгоритмические детали в них не описываются.

Отметим, что производится множество специальных сканеров книг, цифрующих со скоростью 200-400 страниц в минуту. Однако для ветхих исторических документов автоматическое перелистывание не годится из-за возможности их повреждения.

Можно выделить следующие уровни цифровой обработки изображений документов:

- геометрическая коррекция (в процессе сканирования форма оригинала документа может быть искажена, рис. 2),
- яркостная коррекция (рис. 2);
- тематическая обработка первого уровня – ориентация на создание псевдо-гипертекстового документа подобно оцифрованным книгам, представленным на сайте http://books.google.com/; результатом является представление документа в формате «текст в графическом виде» без коррекции (рис. 3);



5. Folgerungen aus den Capillaritätserscheinungen; von Albert Einstein.

Bezeichnen wir mit γ diejenige Menge mechanischer Arbeit, welche wir der Flüssigkeit zuführen müssen, um die freie Oberfläche um die Einheit zu vergrössern, so ist γ nicht etwa die gesamte Energiezunahme des Systems, wie folgender Kreisprocess lehrt. Sei eine bestimmte Flüssigkeitsmenge vorhegend von der (absoluten) Temperatur T_1 und der Oberfläche O_1. Wir vermehren nun isothermisch die Oberfläche O_1 auf O_2, erhöhen die Temperatur auf T_2 (bei constanter Oberfläche), vermindern dann die Oberfläche auf O_1 und kühlen dann die Flüssigkeit wieder auf T_1 ab. Nimmt man nun an, dass dem Körper ausser der ihm vermöge seiner specifischen Wärme zukommenden keine andere Wärmemenge zugeführt wird, so ist bei dem Kreisprocess die Summe der dem Körper zugeführten Wärme gleich der Summe der ihm entnommenen. Es muss also nach dem Princip von der Erhaltung der Energie auch die Summe der zugeführten mechanischen Arbeiten gleich Null sein.

Es gilt also die Gleichung:

$$(O_2 - O_1)\gamma_1 - (O_2 - O_1)\gamma_2 = 0 \quad \text{oder} \quad \gamma_1 = \gamma_2 .$$

Dies widerspricht aber der Erfahrung.

Es bleibt also nichts anderes übrig als anzunehmen, dass mit der Aenderung der Oberfläche auch ein Austausch der Wärme verbunden sei, und dass der Oberfläche eine eigene specifische Wärme zukomme. Bezeichnen wir also mit U die Energie, mit S die Entropie der Oberflächeneinheit der Flüssigkeit, mit s die specifische Wärme der Oberfläche, mit w_0 die zur Bildung der Oberflächeneinheit erforderliche Wärme in mechanischem Maass, so sind die Grössen:

$$dU = s \cdot O \cdot dT + (\gamma + w_0)\, dO$$

und

$$dS = \frac{s \cdot O \cdot dT}{T} + \frac{w_0}{T}\, dO$$

vollständige Differentiale. Es gelten also die Gleichungen:

Первая страница статьи А. Эйнштейна «Следствия из явлений капиллярности»

Рис. 3. Книга Google А. Эйнштейн Работы по кинетической теории, теории излучения и основам квантовой механики. М.: Наука, 1966. Текст представлен изображениями без реставрации

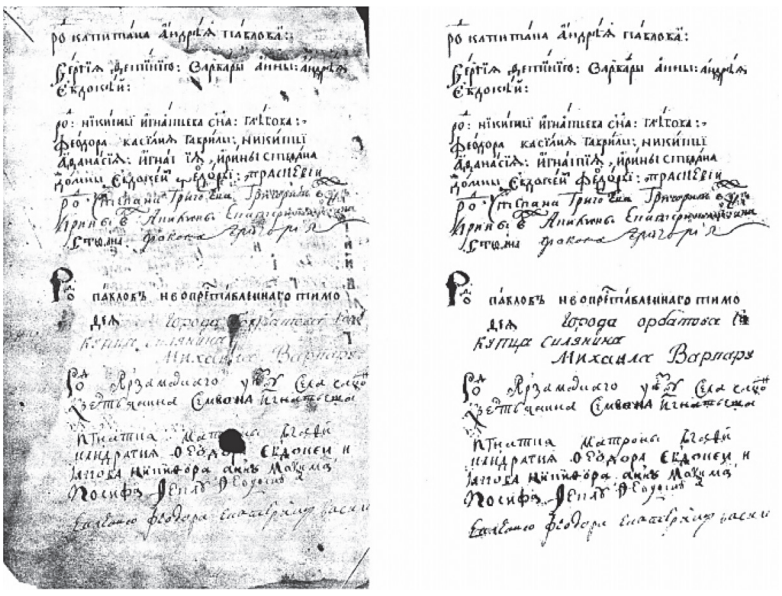


Рис 4. Пример реставрации старого текстового документа [7]: *а* – до реставрации; *б* – после реставрации

• тематическая обработка второго уровня – ориентация на визуальный анализ изображений документов (редактирование изображений, корректировка фона, улучшение изображений, выделение дефектных областей и определение типов дефектов, исправление дефектов, форматирование документа, создание электронной книги и т. п.); результатом является изображение страницы документа или ее части (на рис. 4 приведен пример реставрации документа группой

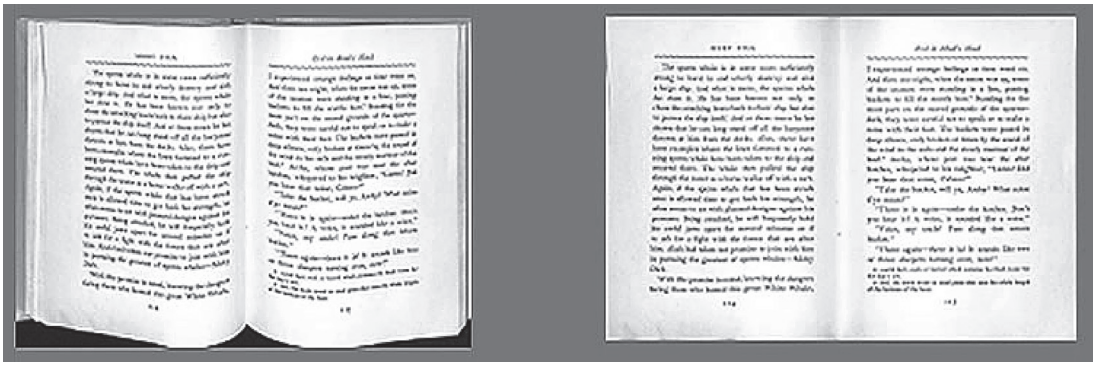


Рис 5. Пример геометрической коррекции отсканированной страницы

российских ученых, но в настоящее время они прекратили разработки в этом направлении);

• тематическая обработка третьего уровня – ориентация на автоматизированный анализ изображений документов и извлечение информации (сегментация на текст+фон+иллюстрации, подобно алгоритму DjVu), утоньшение текстовых символов для автоматического распознавания, повышение качества текста; результатом является комбинированное представление документа в виде «текст»+«изображение текста» с возможностью использования достоверного графического представления документа и работы с текстом (чтение, копирование, поиск).

Алгоритмы цифровой реставрации изображений документов можно разделить на следующие группы:

• автоматические и интерактивные,

• низкоуровневые и контекстно-ориентированные (высокоуровневые).

Рассмотрим основные задачи цифровой реставрации исторических документов:

• формирование цифрового представления документа (сканирование или фотографирование),

• геометрическая коррекция страницы (автоматическая обрезка, исправление перекоса, поворот, выравнивание строк, см. рис. 5),

• исправление неравномерного фона,

• исправление баланса белого,

• фильтрация изображения документа,

• устранение проступающих с обратной стороны надписей,

• выравнивание яркости и контраста,

• преобразование исходного изображения документа в полутоновое и черно-белое представление,

• сегментация изображения на текстовые блоки, иллюстрации и фон,

• заполнение царапин, надрывов и дыр на цифровом представлении страницы,

• детекция пятен и дыр, формирование их маски,

• адаптивная бинаризация текста,

• усиление контраста исходного представления текста,

• формирование изображения однородного представления текстуры основы (бумаги),

• распознавание печатных символов определенного алфавита.

Заключение

В настоящее время электронные библиотеки содержат два основных типа книг: 1) книги, набранные на компьютере и сохраненные в формате *pdf* или *djvu*, 2) книги оцифрованные с помощью сканера и фотокамеры и представленные набором изображений без обработки.

Впервые в Беларуси рассматривается актуальная проблема цифровой реставрации отсканированных или сфотографированных исторически значимых текстовых документов. Показано, что задачи традиционной реставрации таких документов в основном ориентированы на физическое сохранение их носителей (в основном бумажных). При этом основным принципом является «не навреди оригиналу», т. е. носителю информации.

После оцифровки, оригинальный документ испортить невозможно, он остается в хранилище. Цифровая реставрация позволяет менять и обрабатывать электронные копии оригинальных документов без ущерба подлинникам, но с ориентацией на различные применения: обеспечить доступ к зафиксированной на них информации без визуального изменения исходного цифрового представления документа, улучшение яркостных характеристик и геометрическая коррекция представления текстовых строк, выделение исправление и дефектов, преобразование документа в форму электронной книги, формирование комбинированного

представления документа в форме «текст» + «изображение текста», совмещающего достоверное графическое представление исходного документа и возможность работы с его текстом путем копирования, индексирования, поиска.

Создание цифровых экземпляров исторических документов и старых книг позволит библиотекам и архивам создавать цифровые версии оригинальных документов и расширить круг пользователей, некоторые материалы могут быть доступны дистанционно.

Литература

1. Электронный ресурс. Крупнейшие цифровые библиотеки мира. Справка http://ria.ru/society/20090527/172428361.html. Дата доступа 6.04.2015.

2. Электронный ресурс. About the european library services for libraries. http://www.theeuropeanlibrary.org/tel4/aboutus. Дата доступа 6.04.2015.

3. Электронный ресурс. Старопечатные книги. http://elibrary.rsl.ru/?menu=s410/elibrary/elibrary4454/elibrary44544455/&lang=ru. Дата доступа 6.04.2015.

4. Электронный ресурс. Проекты по оцифровке Владимирской областной научной библиотеки. http://library.vladimir.ru/proekty-po-ocifrovke-vladimirskoj-oblastnoj-nauchnoj-biblioteki.htm Дата доступа 6.04.2015.

5. **Носевич В.** Как сберечь цифровое наследие // Архівы і справаводства. 2011. № 6 (78). С. 82–92.

6. Реставрация документов на бумажных носителях: методические рекомендации / Главархив. ВНИИДАД. –. М., 1989. – 152 с.

7. **Канунова Е. Е., Орлов А. А., Садыков С. С.** Методы и алгоритмы реставрации изображений архивных текстовых документов. – М. : Мир, 2006. – 135 с.

8. **Садыков С. С., Канунова Е. Е., Варламов А. Д.** Автоматизированная реставрация изображений архивных текстовых и фотографических документов // Автоматизация и современные технологии. – 2007. – № 8. – С.10–15.

9. **Караваев В. С.** Оцифровка архивных документов: технические и технологические проблемы // Документ. Архив. История. Современность. 2014. – Вып. 14. – С. 243–257.

10. **Бокштейн И. М., Кузнецов Н. А., Мерзляков Н. С., Рубанов Л. И.** Возможности и средства цифровой реставрации архивных рукописных текстов // Информационные технологии и вычислительные системы, М.: ИВВС РАН, 1997. – № 1. – С. 1–15.