Generated by Foxit PDF Creator © Foxit Software
http://www.foxitsoftware.com For evaluation only.

Автоматическая оцифровка книг

В.В. Прудникова,
ДЦмаг-6-1

Последние десятилетия XX века характеризуются быстрым совершенствованием и развитием электроники и компьютерных информационных технологий. Именно в этот период практически все издательства перешли на компьютерный набор и верстку газет, журналов и книг. Издание хранится в памяти компьютера все время набора и верстки, т. е. остается в электронной (невещественной) форме в течение всего процесса подготовки, вплоть до вывода на принтер так называемого постраничного оригинал-макета. Полностью сверстанное и подготовленное к печати издание, хранимое в памяти компьютера или в специальном запоминающем устройстве долговременного типа, можно назвать «электронным изданием».

Электронные издания стали средством комплексного информационного воздействия на человека, сравнимым с радио, кино и телевидением, а в чем-то даже превосходящим эти важные средства массовой коммуникации.

Принципиальным отличием печатных от электронных изданий является возможность интерактивной реализации последних, при которой пользователь может не только перемещаться по встроенным в текст гиперссылкам, но и активно вмешиваться.

Я выделяю основные преимущества электронных книг:

1) сканирование книг и распознавание книг поможет Вам сохранить место в библиотеках и Вашем доме;

2) отсканированную книгу легко передать друзьям;

3) многие мобильные устройства позволят Вам держать всю Вашу отсканированную библиотеку в кармане;

4) старинные и редкие книги благодаря сканированию никогда не пропадут, не сгорят, не исчезнут.

На рис. 1 изображен профессиональный сканер ЭЛАР План-Скан Серии «С» для оцифровки толстых книг, объемных предметов размером до формата А2+, в том числе альбомов, чертежей, карт.

Generated by Foxit PDF Creator © Foxit Software
http://www.foxitsoftware.com For evaluation only.

Рис. 1. Профессиональный сканер для сканирования книг

«Электронные книги» — это хранимый в компьютере текст, оформленный в виде, свойственном печатным книгам. Так, электронные книги обычно дробят содержащий текст на равномерные нумерованные страницы. Их типографика соответствует уровню печатных изданий. Важно сразу же различать сканированные и верстанные электронные книги.

Сверстанные книги — это либо материал, подготовленный авторами в издательской системе, либо распознанная и вручную вычитанная и отформатированная бумажная книга. Оригиналом является обычно файл в формате наглядного текстового процессора (OpenOffice.org или Word) или на языке разметки (LaTeX). Конечным результатом является электронная книга в формате PDF (Adobe Portable Document Format), реже ПостСкрипт (Adobe PostScript) или DVI. Такие файлы обычно содержат векторные шрифты и иллюстрации высокого качества, поэтому они пригодны для печати в любом разрешении, для просмотра на экране, и для поиска по тексту книги (включая возможность выделять и копировать фрагменты текста и иллюстрации). Файлы этого вида кратко будем называть векторными. Типичные векторные PDF-файлы имеют размеры от 3 (редко) до 10–15 килобайт на страницу, в зависимости от числа формул и иллюстраций.

Сканированные книги — это файлы, хранящие целые электронные изображения каждой страницы книги. Такие файлы делаются путем сканирования бумажной книги постранично и дальнейшей обработки с целью улучшения качества и уменьшения размеров файла. По-

Generated by Foxit PDF Creator © Foxit Software
http://www.foxitsoftware.com For evaluation only.

скольку каждая страница хранится в виде ряда точек (растра), то такие книги можно называть растровыми. Основные форматы, употребляемые для растровых файлов, это PDF и DJVU. В этих форматах можно добавить также и распознанный текст, закладки и гиперссылки, чтобы были возможны быстрые переходы по книге и автоматический поиск текста. Поэтому качественно сделанные растровые книги не менее удобны в использовании, чем векторные, и несущественно проигрывают им в качестве распечатанного текста. Типичный размер растровой книги — от 5 (редко) до 10–15 килобайт на страницу, в зависимости от разрешения и качества текста или иллюстраций.

Производство векторной электронной книги из бумажной книги путем компьютерного распознавания (OCR) связано с колоссальными затратами труда, особенно если книга содержит много иллюстраций, графиков, диаграмм, таблиц, или формул, ибо нынешнее состояние распознавательных программ заставляет форматировать все это вручную, и нередко — исправлять ошибки распознавания текста. Поэтому для таких книг гораздо легче делать именно растровые, а не векторные электронные версии. Даже в чисто текстовых книгах — без иллюстраций, таблиц или формул — автоматическое распознавание порой дает трудно-выявляемые ошибки. Гораздо быстрее приготовить растровую электронную книгу, тем более что современная технология сжатия изображений позволяет делать файлы вполне приемлемого размера. Например, средний размер растровых книг, включая распознанный текст — 13 КБ на страницу. Это означает, что растровая книга, имеющая 400 страниц, в среднем занимает около 5 МБ (цифры приводятся для формата DJVU). На стандартном DVD-носителе (4,3 ГБ) может поместиться около 900 таких книг.

Некоторые программы позволяют делать файлы формата PDF, в которых весь плохо распознанный материал содержится в виде отсканированных картинок, а текст является векторным. Такие PDF-файлы, однако, сильно проигрывают чисто растровым книгам и по внешнему виду (нестыковка векторных шрифтов и фрагментов изображения страницы), и по размеру файлов.

Основные моменты сканирования

Есть два основных метода сканирования: целым разворотом и по одной странице. При любом методе части страницы, где находится текст, должны быть полностью прижаты к стеклу — иначе возникает характерное затемнение в области корешка из-за наклонного падения света лампы подсветки (в любых сканерах) и размывание текста из-за малой глубины резкости (только в CIS-сканерах). Поэтому практически невозможно хорошо отсканировать книги, которые плохо открываются. Существует, однако, специальный сканер OpticBook 3600, позволяющий сканировать книгу, открытую на 90 градусов. Этот сканер пока имеет

Generated by Foxit PDF Creator © Foxit Software
http://www.foxitsoftware.com For evaluation only.

драйверы только под Windows. При сканировании разворотом, если книга полностью не помещается на стекло, может возникать дополнительное размытие текста на краю страницы из-за того, что текст не прижат к стеклу. Все это необходимо тестировать перед началом сканирования и определить оптимальный метод.

Необходимо также определить оптимальную рамку сканирования и контрастность. Это можно подобрать только методом проб и ошибок, но это необходимо подобрать правильно перед сканированием. Желательно также сразу же установить гамму и точки белого и черного, если это позволяет софт Вашего сканера.

Класть книгу абсолютно ровно, без перекосов, у Вас все равно не получится, но это и не нужно, так как выравнивание можно проделать автоматически специальным софтом на стадии обработки сканов. Разрезание отсканированных разворотов и выравнивание полей тоже делается почти автоматически.

Иногда имеет смысл сделать ксерокопию с книги и сканировать этот ксерокс. Дело в том, что в ксерокс-машинах гораздо больше глубина резкости, чем у сканеров, и поэтому удается скопировать участки текста, прилегающие к корешку книги, даже если книга плохо раскрывается. Кроме того, ксероксы специально делают повышение контрастности изображения, что позволяет избавиться от серого фона страниц и максимально уменьшить затемнение в середине разворота.

Софт для управления сканером можно использовать любой, например VueScan (есть для Windows / Linux), Irfan View (Windows), XnView (Windows), gimp/sane (Linux). Поддержка сканеров для Linux/Mac гораздо хуже, потому что, как правило, все производители делают драйверы только для Windows.

Я не советую пользоваться программой FineReader для сканирования текста при создании электронных книг, потому что FineReader автоматически делает неаккуратное выпрямление косых сканов, из-за которого в изображении появляются «изломы». Эти изломы не мешают при распознавании текстов, но плохо выглядят в растровом варианте отсканированной книги. Пример отсканированного текста показан на рис. 2. На тексте появились «изломы» из-за сканирования в FineReader.

Владелец
Важнейшей
предпринимательство

Рис. 2. Отсканированный текст с видимыми изломами

Generated by Foxit PDF Creator © Foxit Software
http://www.foxitsoftware.com For evaluation only.

Софт для сканирования желательно использовать такой, чтобы получить сырые сканы в формате TIFF (не JPG, так как неизбежна потеря качества), поименованные автоматически, например, так: page0001.tiff, page0002.tiff и так далее.

При сканировании по одной странице бывает удобнее сканировать сначала все четные страницы, потом все нечетные, и только потом переименовать все файлы автоматически по возрастанию номеров. При сканировании из XnView можно сразу указать, что сканируются только четные или только нечетные страницы, — номера будут проставляться автоматически.

Разрешение (цифра «dpi» — количество пикселей на дюйм) и глубина цвета (черно-белый — 1 бит, серый — 8 бит, цветной — 24 бит) — самые важные параметры сканирования. Софт для сканера должен давать пользователю возможность выбрать эти параметры непосредственно, а не просто выбирать между непонятными режимами типа «текст» — «рисунок» — «фото для Интернета». Сканер должен поддерживать разрешение как минимум 600 dpi. Сканировать можно либо в 600 dpi, либо в 300 dpi. Я не рекомендую сканировать издание в 150 или 200 точек на дюйм, т. к. время при этом мы сэкономить не сможем, а качество будет безвозвратно утеряно. Другие разрешения, такие как 360, 400 и т. д., мало смысла использовать — они работают через интерполяцию изображения, то если реально сканер сканирует в 300 dpi в сером и делает интерполяцию до 400 dpi черно-белого. Такую интерполяцию можно и нужно сделать специальным софтом, а не тем софтом, что пришел со сканером.

Многие сканеры одинаково быстро делают скан листа в 300 dpi черно-белого режима и в 300 dpi серого режима (greyscale). Поэтому если вы сканируете в 300 dpi (а не в 600 dpi), то лучше сканировать всегда в сером режиме (greyscale), даже если книга не содержит вообще цветного материала. Специальный софт потом может поднять разрешение 300 dpi серых сканов до 600 dpi черно-белого, при сохранении отличного качества текста (как будто сканировали в 600 dpi черно-белом). Проблемы 300 dpi серых сканов видны только в литературе с фотографиями, передаваемыми растровым методом: в некоторых местах появляется эффект муара, то есть волн яркости и/или цвета по изображению.

Для максимального качества рекомендуется книги сканировать в сером режиме при 600 dpi, если же в книге есть информативные цветные иллюстрации, то в цветном 600 dpi. Это разрешение позволяет полностью разбить растровое изображение на отдельные цветные точки, что полностью подавляет муар.

Сканирование цветных материалов

Сканирование и обработка цветных материалов (книги с большим количеством фотографий, важных для содержания) связаны

Generated by Foxit PDF Creator © Foxit Software
http://www.foxitsoftware.com For evaluation only.

с существенно большими трудностями, чем сканирование черно-белых книг. Гораздо труднее получить конечный файл разумного размера при сохранении хорошего качества изображения. Исходные отсканированные файлы могут достигать ста мегабайт и больше на страницу, а конечный результат — сотен килобайт на страницу.

Обработка отсканированных изображений

После сканирования необходимо просмотреть все страницы и убедиться, что нет явных огрехов. Например, иногда по недосмотру книга неровно легла на стекло сканера и часть текста на какой-либо странице не отсканирована, или были вовсе пропущены некоторые страницы. После этого можно архивировать отсканированные изображения и приступать к обработке. Поскольку сканирование — физически самый трудоемкий этап, рекомендуется держать резервную копию всех исходных сканов (такими, какими они были до обработки) на случай какого-либо сбоя.

Каковы главные задачи обработки? Они зависят от того, ставим ли мы целью создание векторного файла или растрового файла. Для создания векторного файла производится распознавание (OCR) текста и его дальнейшее редактирование вручную в текстовом процессоре (таком, как MS Word или Adobe Pagemaker). Конечным продуктом обычно является сверстанная книга в формате PDF. Для создания растрового файла необходима доводка графических изображений до высокой степени сжатия и качества, а распознавание (OCR) производится лишь начерно, без вычитки и правки текста, в самом конце процесса. Обработка графических изображений производится обычно в пакетном режиме, так что не требуется обрабатывать каждую страницу вручную в Photoshop или другом графическом редакторе. Поэтому затраты времени на создание растровой электронной книги гораздо меньше, чем на создание векторной книги.

Графическая обработка сканов состоит из следующих основных шагов:

- преобразование серых сканов в черно-белые (если исходные сканы были серыми в 300 dpi, то после этого получаются черно-белые в 600 dpi);
- разрезание разворотов на два изображения отдельных страниц (если книгу сканировали в развороте);
- поворот изображения каждой страницы, чтобы текст стал, по возможности, горизонтальным;
- отрезание ненужных темных полос на краях, создание ровных и одинаковых для всех страниц белых полей;
- вычищение «грязи» на страницах (включая помарки от руки, штампы и прочее).

Generated by Foxit PDF Creator © Foxit Software
http://www.foxitsoftware.com For evaluation only.

Эти шаги частично автоматизированы в программе «Scan Kromsator» (Windows) и описаны в инструкции «Scan and Share». Однако если Scan Kromsator показался для вас слишком сложным, вы можете воспользоваться программой Scan Tailor. После создания чистовой версии всех страниц книги, которые пока что хранятся в отдельных графических файлах, приступают к сжатию всех страниц в единый файл формата DJVU или PDF.

Файлы PDF и DJVU могут использовать разные степени сжатия. Наибольшее сжатие достигается в формате DJVU (алгоритм JBIG2), если текст черно-белый, отсканирован четко (это сильно зависит от физического состояния исходной книги), шрифт не слишком мелкий, а края букв ровные (не рваные). Формат PDF позволяет сжимать как алгоритмом JBIG2 (при этом размер получается на 20–30% больше, чем размер DJVU), так и менее эффективными алгоритмами, например TIFF-G4. Размер PDF-файла после сжатия PDF/TIFF-G4 примерно в 4–8 раз больше, чем у PDF/JBIG2.

Имеются программы для создания хорошо сжатых DJVU- и PDF/JBIG2-файлов. Для формата DJVU это коммерческие программы от LizardTech: DjvuSolo и Djvu Document Editor. Для формата PDF это коммерческая версия Adobe Acrobat (не Reader). Есть и бесплатные программы для создания DJVU и PDF/JBIG2, но они пока не дают настолько хорошего сжатия, как коммерческие версии. Программа CPCtool, используемая как промежуточный этап перед окончательным сжатием, позволяет несколько улучшить сжатие DJVU (10–30%) и во многих случаях сгладить «лохматые» контуры букв.

После создания окончательной чистовой версии книги делается распознавание текста (OCR). Распознавание текста на большинстве языков можно производить как коммерческой версией Djvu Document Editor (для DJVU), так и широко распространенной программой FineReader (для PDF). Имеется также бесплатный софт (утилита DjvuOCR) для вставления OCR-слоя в DJVU-файлы после распознавания через FineReader. По опыту, FineReader дает лучшее качество распознавания, чем Djvu Document Editor (который использует движок IRIS).

Имеется также возможность автоматически добавить гипертекстовые ссылки в оглавление и индекс DJVU-книги. Это делает бесплатная утилита Djvu Hyperlink Editor.

Часто бывает необходимо улучшить уже имеющуюся электронную книгу. Доделка бывает по разным причинам необходима как для сверстанных, так и для сканированных книг. Поскольку сканирование или верстка — самый трудоемкий этап, то целесообразно обработать уже имеющийся файл до максимально хорошего качества (за исключением крайних случаев, когда качество имеющегося файла книги слишком низкое и лучше переделать все заново).

Generated by Foxit PDF Creator © Foxit Software
http://www.foxitsoftware.com For evaluation only.

Методики оцифровки

В прошлом применялся ручной набор текста книги.

Сегодня процесс оцифровки включает два подхода.

1. Обязательный: получение копий страниц в виде графических (обычно растровых) изображений, осуществляемый путем сканирования или фотографирования с последующей обработкой и сохранением в одном из форматов графических файлов. В этом случае полностью сохраняется оригинальная верстка книги, и исключаются какие-либо ошибки, однако невозможен поиск или извлечение фрагментов текста для, например, целей цитирования.

2. Опциональный: распознавание текста (технология «оптического распознавания символов» — OCR) с последующим сохранением распознанного текста в одном из форматов электронных книг. В этом случае становится возможен полнотекстовый поиск по книге и индексация больших массивов электронных книг, однако затрудняется воспроизведение оригинальной верстки, изображений, схем и формул, практически неизбежными становятся ошибки распознавания.

В последнее время (особенно с появлением формата DjVu) все чаще применяется смешанный подход: текст книги распознается в автоматическом режиме и подкладывается под оригинальные растровые изображения страниц, что позволяет совместить преимущества обоих подходов.

Что такое ArcScan for ArcGIS?

ArcScan — дополнительный модуль для ArcGIS, разработанный для преобразования растровых данных в векторные. Этот простой в использовании продукт представляет собой набор мощных команд и инструментов для оцифровки бумажных карт.

Благодаря тому, что этот модуль полностью интегрирован в среду ArcGIS, имеется возможность задавать топологические правила для векторизуемых слоев, и работать с моделями данных, разработанными в ArcGIS, поддерживая, таким образом, целостность данных уже на этапе оцифровки.

При работе с модулем можно использовать все предоставляемые ArcMap возможности для редактирования растровых и векторных данных.

Начиная с ArcGIS версии 9.1, ArcScan уже включен в ArcEditor и ArcInfo, так что дополнительно приобретать его нужно лишь при работе с ArcView.

Используя ArcScan возможно:

- создавать линейные и полигональные векторные объекты в форматах базы непосредственно по растровому изображению;

Generated by Foxit PDF Creator © Foxit Software
http://www.foxitsoftware.com For evaluation only.

- векторизовать объекты (переводить из растрового формата в векторный) в интерактивном или автоматическом режимах;
- подготавливать (очищать) изображения для векторизации в автоматическом режиме;
- задавать среду замыкания для растров;
- выбирать группы ячеек растров путем запроса к связанным с ним областям.

ArcScan позволяет выполнять векторизацию в трех режимах:

- автоматическом (batch mode);
- полуавтоматическом или интерактивном (tracing);
- ручном (head-up digitizing).

Автоматическая векторизация существенно сокращает время, затрачиваемое на оцифровку растровых изображений. В этом режиме существуют два способа векторизации: centerline и outline (рис. 3).

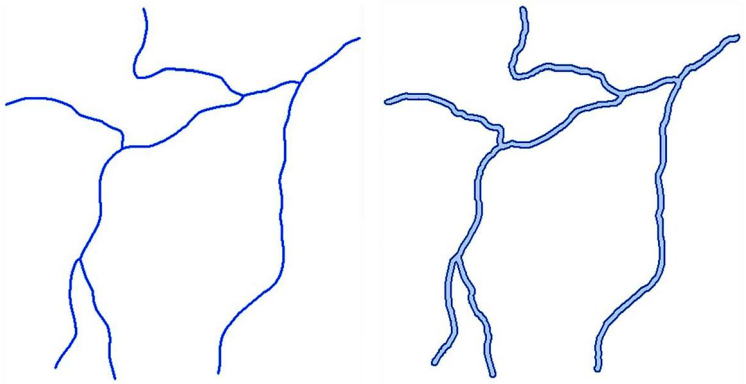

Рис. 3. Centerline (*а*) и Outline (*б*)

В режиме centerline строятся центральные линии растровых линейных объектов и границы площадных объектов.

В режиме outline строятся границы всех растровых связанных компонент в виде полигонов. Результатом является множество полигонов.

Полуавтоматическая или интерактивная векторизация (трассирование) применяется в тех случаях, когда требуется больший контроль над процессом векторизации или нужно векторизовать небольшую часть изображения. Пример трассировки текста приведен на рис. 4.

С помощью курсора задается начальная точка и направление трассирования, после чего автоматически строится центральная линия от начальной точки до конца растровой линии, если по пути не встретится площадной объект или точка пересечения с другой линией. Если центральная линия попадает в точку пересечения, то трассиров-

Generated by Foxit PDF Creator © Foxit Software
http://www.foxitsoftware.com For evaluation only.

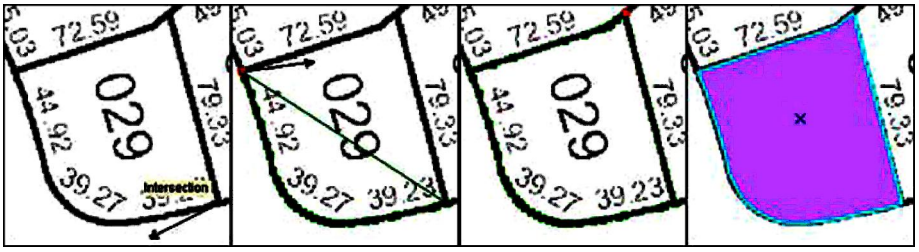

Рис. 4. Пример трассировки текста

щик останавливается и ждет, пока оператор снова укажет направление, в котором нужно продолжить трассирование. Если центральная линия остановилась на границе площадного объекта, то оператор должен перейти в режим ручного цифрования и оцифровать этот объект.

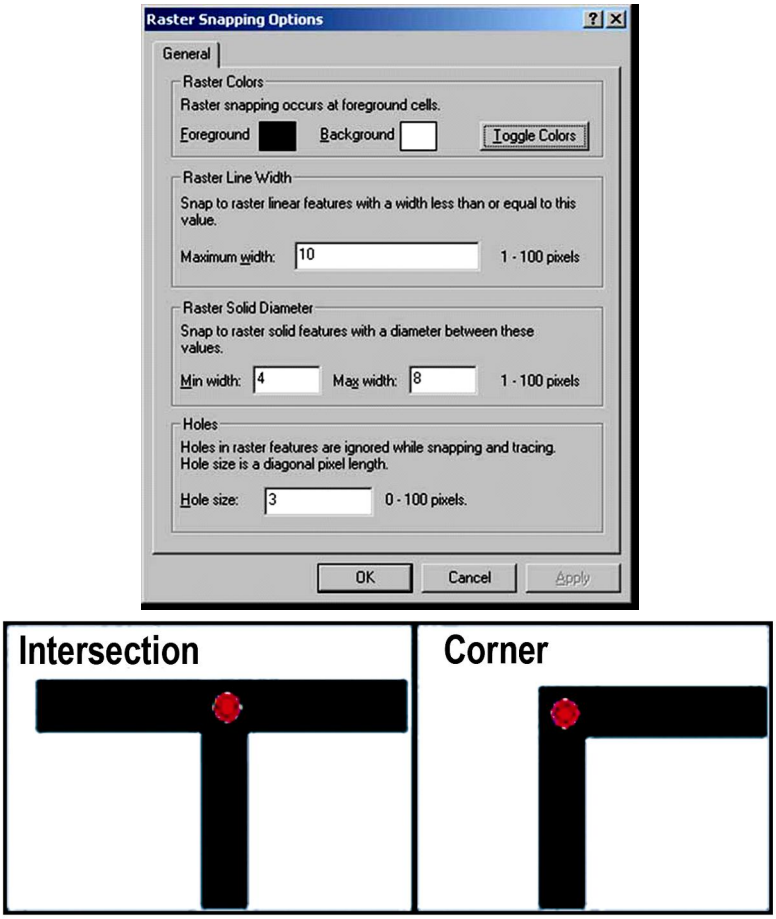

Рис. 5. Установка опций замыкания для растра и примеры замыкания

Generated by Foxit PDF Creator © Foxit Software
http://www.foxitsoftware.com For evaluation only.

Ручная оцифровка позволяет оператору осуществлять непрерывный контроль над процессом векторизации, задавая с помощью курсора положение каждой вершины центральной линии. Ручная оцифровка используется для векторизации растровых изображений плохого качества, изображений, содержащих сразу нескольких тематических слоев, а также при наличии сложных видов линий. Ее также целесообразно использовать для оцифровки прямых линий.

В данном режиме есть особый инструмент – raster snapping, позволяющий автоматически привязывать начальную точку к центральной линии, точке пересечения линий, концам линий или углам. Быстрое наведение курсора на специфические точки повышает точность и увеличивает эффективность оцифровки, так как отпадает необходимость в частом изменении масштаба изображения на экране.

ArcScan также можно использовать для редактирования черно-белого растрового изображения до начала векторизации. Для этого имеются стандартные инструменты (Brush, Fill, Line, Erase), обычно используемые в других векторных редакторах. Есть еще два инструмента: Swap – для инвертирования цвета фона и объектов, и Magic Erase – для удаления целиком связанной компоненты.

Имеются также инструменты, сочетание которых удобно для автоматического удаления изолированных шумов (белых и черных пятен) с растрового изображения перед автоматической векторизацией (рис. 6).

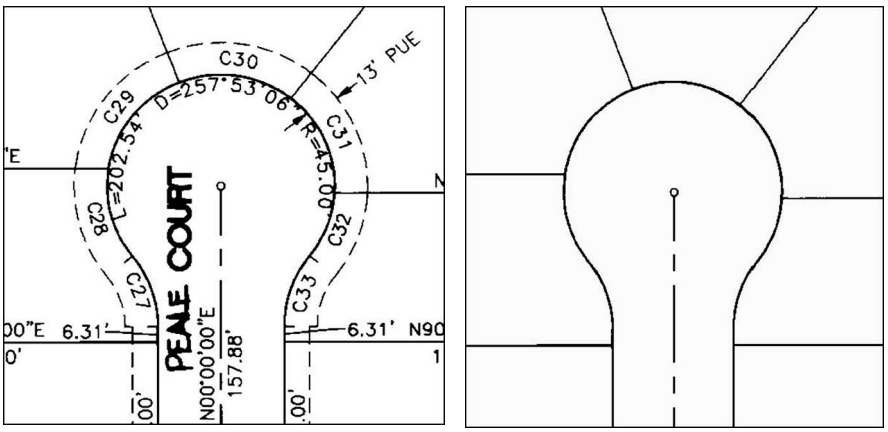

Рис. 6. До и после очистки растра

Когда результат получен, то ArcScan позволяет провести векторную постобработку – сгладить и, затем, генерализовать полученные в результате векторизации линии и границы полигонов. Важно, что в используемом алгоритме интенсивность сглаживания и генерализации не зависят от толщины линии.

Generated by Foxit PDF Creator © Foxit Software
http://www.foxitsoftware.com For evaluation only.

Центральные линии могут быть прерывистыми из-за использования сложных условных знаков (штриховые линии и т. п.) и других помех. Одной из операций векторной постобработки является автоматическое замыкание пробелов. Этот алгоритм использует два параметра: максимальную длину пробела и угол, внутри которого может лежать продолжение (рис. 7).

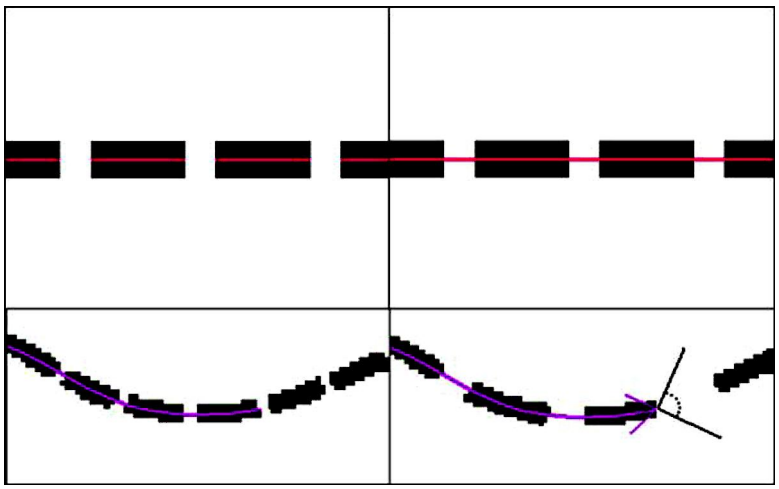

Рис. 7. Параметры векторизации

Настройка: выбор порогов и управляющих параметров производится в интерактивном режиме. Оператор меняет значения параметров и сразу же видит результаты сглаживания, генерализации и замыкания пробелов в режиме предварительного просмотра видимой части изображения. Подобрав необходимые параметры, можно запустить режим автоматической или полуавтоматической векторизации.

Однажды выбрав оптимальные параметры для векторизации карт определенного типа, их можно сохранить в отдельный стиль, и использовать в дальнейшем.

Интеграция полиграфической продукции с электронными документами приносит чисто практические выгоды. Так, переход на цифровую форму позволяет обеспечить сохранность многих уникальных видов продукции, как например рукописи. Даже обычные фотографии и картины теряют со временем свои качества. Хранение их электронных копий позволяет донести до последующих поколений уникальные культурные шедевры. Наконец, хранение документов и изданий в электронной форме позволяет организовать электронные базы данных, четкая структура и развитые средства поиска и навигации в которых облегчают процесс обнаружения нужных материалов и их фрагментов.

Generated by Foxit PDF Creator © Foxit Software
http://www.foxitsoftware.com For evaluation only.

Библиографический список

1. *Бондаренко А.В.* Исследование подходов к построению систем автоматического считывания символьной информации / А.В. Бондаренко, В.А. Галактионов, В.И. Горемычкин, А.В. Ермаков, С.Ю. Желтов. – М., 2003.

2. *Кузнецов А.Б.* Роль современных информационных технологий в поддержке и развитии культурной компетентности / А.Б. Кузнецов. – М., 2008.

3. *Масевич А.Ц.* Новые технологии в информационном обеспечении науки / А.Ц. Масевич, Е.А. Савельева, А.К. Багажков. – М., 2009.

4. Ресурсы Интернета – http://portal.lgo.ru;

5. Ресурсы Интернета – http://www.lib.scu.ru;

6. Ресурсы Интернета – http://www.publish.ru;

7. Ресурсы Интернета – http://www.videoscan.msk.ru;

8. Ресурсы Интернета – http://www.yandex.ru.