В. Б. БОВИН

ОПЫТ СКАНИРОВАНИЯ ФОЛЬКЛОРНОГО РУКОПИСНОГО АРХИВА ИЯЛИ КарНЦ РАН[1]

В рамках проекта по созданию информационной системы фольклорного архива Института языка, литературы и истории КарНЦ РАН были выполнены работы по сканированию некоторой части рукописных документов научного архива. Целью работы было создание архива нового поколения, существенно упрощающего процессы хранения, поиска и работы с информацией. В связи с этим потребовалась выработка определенной методики, связанной с техникой сканирования и хранения информации. Наличие таких документов в электронном виде предоставляет быстрый поиск и обеспечивает удобную для исследователя работу с материалами архива. Копирование электронных документов происходит достаточно быстро и без потерь качества. Отсканированные материалы можно отправлять по электронной почте, публиковать в сети Интернет или распечатывать на печатающем устройстве. Например, поиск необходимых документов среди стопки папок внушительных размеров отнимет несколько рабочих дней. Пролистывание электронных документов займет несколько часов работы. Интегрирование электронных документов в базу данных позволит отыскать необходимые исследователю материалы в течение нескольких минут.

Как правило, исходные документы (фольклорные коллекции) представляют собой различной толщины подшивки, тетради с записями от руки или машинописные листы. Преобразование в электронный вид такого документа подразумевает создание электронной копии высокого качества, чтобы при печати или при издании документа изображение было бы не хуже оригинала. Существует два вида электронного представления такого рода информации: графическое и текстовое.

Под графическим видом представления понимают перевод в электронную форму всего изображения документа. Это можно сравнить с микрофильмированием книг, которое выполнялось в советское время. Например, в Российской национальной библиотеке в таком виде хранятся «Народные произведения Олонецкого края. Вып. 11. Стихотворения Д. А. Мошинского 1918 г.».[2] Документ представляет из себя специальную, позитивную (слайдовую) фотопленку с кадрами всех страниц книги.

Текстовое же представление лишь по содержанию повторяет исходный документ, визуально же не имеет ничего общего с первоисточником. Например, при подготовке к печати словаря поморского языка И. М. Дурова преследовались две цели. Первая — компьютерный набор текста словаря. Вторая — сохранение оригинала рукописи в электронном виде. Для целей набора качество картинки должно обеспечить лишь хорошую читаемость букв и слов. Если же потребуется печатать фотографии исходных тетрадей, то такого качества уже недостаточно.

[1] Исследование выполнено при финансовой поддержке Российского гуманитарного научного фонда в рамках проекта «Создание информационной системы по фольклорному рукописному архиву ИЯЛИ КарНЦ РАН», проект № 08-04-12144в.

[2] РНБ. Л30-В-8/20 | 37.59.6.89.

При планировании работ по сканированию оригиналов рукописного архива ИЯЛИ процесс выполнения представлялся четким и ясным. Единицы хранения (листы тетрадей) пропускаются через сканер. Затем полученные файлы собираются в каталогах жесткого диска компьютера. Наконец графические файлы записываются на электронные носители, например DVD+R. Однако позже возникли вопросы, которые потребовалось решать уже в процессе выполнения проекта. По причине того что страницы тетрадей имеют разный размер, автоматически изменяются и расчетные цифры конечного объема электронного архива. Та же картина возникает в случае, если в тетрадях листы заполнены с обеих сторон. Некоторые страницы, подлежащие сканированию, со временем обветшали, и сейчас они уже не такие ровные и гладкие, а это влечет за собой дополнительное время на подготовку документов к сканированию. Тетрадь в отсканированном виде может потребовать более одного DVD+R диска для размещения электронных страниц, в результате чего количество «болванок» DVD+R возрастет. Сканирование одной страницы при высоком разрешении увеличивает время преобразования документа в электронный вид. Наконец при многодневной, многочасовой и монотонной работе неизбежны пользовательские, программные и аппаратные ошибки, поэтому стоит учитывать необходимость обслуживания техники и программного обеспечения. В таких условиях нужно проявлять ответственный подход и работать добросовестно, так как документы с годами истлевают, ветшают и работа по сканированию, осуществляемая в наши дни, может оказаться последним шансом сохранить архив.

Сотрудники архива бережно относятся к документам, тем не менее для качественного сканирования тетради рекомендуется расшивать, проделывая это с величайшей осторожностью. Дело в том, что книгу на стекле планшетного сканера всей поверхностью не разместить. В местах изгиба сканирование края листа затруднено и потому будет произведено с потерей резкости. Разумеется, речь идет о машинописных листах XX в., находящихся в хорошем состоянии. Для более ценных рукописей, где расшивание документов недопустимо, используются дорогостоящие планетарные сканеры, оптика которых позволяет сохранять резкость изображения на всем протяжении листа в оговоренных инструкцией пределах.

Сканирование всегда производится в цвете для того, чтобы сохранить подлинный вид документа. Если в дальнейшем цвет не потребуется, электронный документ всегда можно обработать в графическом редакторе, обесцветив. Разрешение (количество точек рисунка по вертикали и горизонтали) для типографской печати обычно составляет 300 dpi (англ. dots per inch — *количество* точек на дюйм). Для архивных целей в зависимости от ценности документа можно использовать значения от 600 до 1200 и даже более точек на дюйм. Однако чем выше значение dpi, тем дольше производится сканирование страницы, тем больше объем графического файла. На практике в зависимости от значения dpi сканирование листа формата A4 может занять от 1 до 6 минут. Хочется отметить, что опираться необходимо не на значение dpi, а на размеры получаемого со сканера изображения (разрешение рисунка). Однако не все сканеры оперируют такими цифрами. Поэтому в результате пробного сканирования и получении информации об изображении можно произвести несложный расчет. Каждая сторона изображения, допустим 3500×4000 точек, делится на значение dpi, предположим 300, и умножается на дюйм (2.54 см). Согласно вышеизложенным данным при печати 300 dpi можно получить качественно отпечатанный лист размером 29.6×33.8 см. При печати этого документа на бумаге с размерами 59.2×67.6 см (т. е. в два раза больше) получившееся изображение будет выглядеть как отсканированное при 150 dpi, что безусловно ухудшит характеристики распечатанного документа. Таким образом, параметры сканирования документа рекомендуется увеличить вдвое. Для указанного образца следует произвести сканирование с размерами сторон 7000×8000 точек и так далее.

В зависимости от возможностей фольклорных архивов, подавляющее большинство которых не имеют ни государственного статуса, ни штата сотрудников, не говоря уже про достойное финансирование, для сканирования можно использовать планшетные сканеры наподобие Mustek ScanExpress A3, Canon CanoScan, Epson Perfection и другие, приемлемые по стоимости. Уровень качества определяется при сканировании специальных настроечных страниц (тестовых мир). При подборе рабочего разрешения (размеров картинки) ориентируются на результаты, полученные после сканирования тестовой ми́ры. Для архива большого объема потребуется высокопроизводительный сканер, способный выдержать многочасовую ежедневную нагрузку. Использование других типов сканеров, например ручных, или же фотографирование качественным цифровым фотоаппаратом не рекомендуется. При закупке компьютера, входящего в состав сканирующего комплекса, необходимо уделить внимание объемам жесткого диска, оперативной памяти и видеопамяти. Во избежание случайных потерь информации и последующего восстановления создание дисковых массивов типа RAID исключается.

В процессе работ по сканированию документов рукописного архива сотрудник укладывал страницу с изображением на стекло сканера таким образом, чтобы строчки текста по возможности были параллельны краям стекла, и производил оцифровку. При сканировании нами использовалась графическая программа IrfanView (http://irfanview.com) и сканирующее программное обеспечение из комплекта сканера. IrfanView позволила выполнить работу быстрее, при этом не внося в изображение ничего лишнего. Для профессионального сканирования приобретается специальное программное обеспечение, соответствующее высокому уровню выполняемых работ.

При сохранении документов выбирается тип файла. Электронные архивные документы рекомендуется хранить в неупакованном виде, например в формате TIFF-Uncompressed (несжатое растровое изображение). При незначительном повреждении носителя (в данном случае диска DVD+R) пропадет лишь часть видимых точек изображения. При применении графических форматов с использованием различных алгоритмов сжатия (JPEG, PNG, TIF-Compressed) такое повреждение губительно практически для большей части всего файла. Однако при использовании страховочного фонда применение этих форматов допускается. Наконец хотелось бы упомянуть о схеме сжатия графики, присущей формату JPEG2000 (http://www.openjpeg.org), PNG и традиционному JPEG, так называемый прогрессивный режим сжатия (Progressive). При сохранении в JPEG указывается параметр Progressive, и результирующий файл уже менее подвержен разного рода повреждениям. Кроме того, JPEG2000 в отличие от традиционного JPEG позволяет сохранять документы без потерь качества, приближаясь к TIFF. Повреждения исходного файла в объеме 20—30 % почти незаметны на глаз. Для традиционного JPEG это в некоторых случаях равнозначно почти полной потере изображения.

Другая сторона вопроса касается быстро меняющейся техники и устаревания форматов. Если формат JPEG2000 устареет и исчезнет из пользования (другие форматы также не являются исключением), то потребуется искать программу, которая «умеет читать» JPEG2000, в противном случае доступ к архиву будет невозможен.

При сканировании документов рукописного архива мы использовали формат JPEG с коэффициентом сжатия 95—100 %, которое обеспечивает оптимальное качество картинки при значениях dpi от 1200 до 2400. Перед началом выполнения проекта производят пробное сканирование документа и последующее сохранение результата в JPEG Progressive (качество высшее — 100 %), для того чтобы выяснить, соответствует ли полученный электронный документ требуемым расчетным размерам (двукратный размер по сравнению с оригиналом).

В электронном архиве графическим файлам даются имена, которые набираются латинскими буквами, что позволяет безошибочно читать эти файлы программами типа СУБД (система управления базами данных) и работать с ними в разных операционных системах, которые обычно ставятся на компьютеры пользовательского класса. На сегодняшний день формирование структуры файлового дерева научного архива ИЯЛИ практически завершено.[3]

После того как архив отсканирован, создается пользовательский фонд (фонд использования). В электронном архиве сотрудники и посетители работают только с фондом использования, основной же фонд «беспокоят» при создании страховочного фонда и при оценке технического состояния носителей информации. При создании фонда использования исходное изображение посредством графического редактора или специальных программ уменьшается до таких размеров, при которых остается возможность свободного прочтения текста. Полученная картинка сохраняется на диске с применением любых графических форматов, использующих высокие алгоритмы сжатия (рекомендуется JPEG). Полученные документы занимают небольшой объем на внешних накопителях, они могут быть использованы при составлении презентаций и публикаций на страницах сети Интернет. Печать пользовательских документов в типографии практически невозможна из-за небольшого значения dpi и низкого качества картинки, что отчасти может служить своеобразной защитой авторского права.

Наряду с преимуществами электронный архив имеет и недостатки. Во-первых, это недолговечность материала, из которого изготовлен диск DVD+R. Кроме того, информация на диске хранится в рабочем слое, который с двух сторон закрыт бесцветными дисками (пластинами). Если заливка лаком стыка двух пластин произведена некачественно, то внутрь диска попадет воздух, кислород, входящий в его состав, может со временем разрушить отражающий слой алюминия. В результате диск перестанет «читаться». Во-вторых, высокая чувствительность DVD+R к механическим повреждениям. Продольные (по ходу лазерного луча) царапины на бесцветной пластине ведут к ошибкам во время считывания информации. В ходе эксперимента в кабинете звукозаписи ИЯЛИ было также установлено, что две-три глубокие, перпендикулярные царапины, оставленные на рабочей поверхности DVD+R, приводят к отказу чтения диска приводом DVD. При этом сами данные были предварительно подготовлены к легкому восстановлению информации, утилитой для проверки и восстановления поврежденных файлов ICEECC (http://ice-graphics.com/ICEECC/IndexR.html).

Быстрое копирование всего диска решает проблемы, связанные с разного рода износами оптических дисков формата DVD+R. Учитывая это, необходимо запланировать резервное сохранение всего электронного архива через утвержденное количество лет. Из-за того что выполнить такую задачу за один-два дня невозможно (в зависимости от объема архива), для электронного архива имеется специальная тетрадь, в которой отмечены даты записи дисков. Например, через 15—30 лет необходимо создать дополнительный страховочный фонд путем копирования всех дисков архива. Каждый год диски проверяются на чтение. В зависимости от объема проверяется либо весь архив, либо по нескольку дисков из всех коллекций. На основании состояния страховочного фонда можно планировать дату следующего резервного сохранения. Оригинальные документы выбрасывать недопустимо, поскольку в случае утраты электронных носителей основного фонда (при отсутствии страховочного) рукописи потребуется повторно сканировать.

Например, в Фонограммархиве ИЯЛИ КарНЦ планируется следующая методика работы с электронным рукописным фондом. Отсканированные доку-

---

[3] Образцы имен файлов и каталогов для отсканированных документов размещены на сайте http://rst.krc.karelia.ru.

менты записываются на DVD+R, при этом оригиналы с жесткого диска не удаляются. По истечении 6 месяцев все диски DVD+R проверяются на предмет чтения документов специальной программой (например, приложение Nero DiscSpeed). В случае успешного чтения создается страховочный фонд в том же объеме. Диски DVD+R, вышедшие из строя, изымаются из архива и документы записываются повторно (источник — жесткий диск), при этом в тетради ставится соответствующая отметка. Через полгода проверяются уже оба фонда. Как правило, лазерные диски приходят в негодность в течение первых трех-шести месяцев использования. В 2011 г. при довольно низких ценах на жесткие диски было принято решение не удалять электронные документы, применяя жесткий диск в качестве фонда использования, а также источника информации для базы данных рукописного архива.

Методики, рекомендации и форматы, используемые для хранения документов, важны не только при планировании и выполнении работ по сканированию бумажных документов. Готовый электронный архив необходимо грамотно использовать, преподнести пользователю-исследователю, подготовить материалы для публикации. Полученные сведения заметно облегчают работу, систематизируют данные и позволят сохранить архив для будущих поколений исследователей. С рекомендациями по технике сканирования можно ознакомиться в сети Интернет по адресу http://rst.krc.karelia.ru. На сайте освещаются дополнительные вопросы, требующие детального рассмотрения: проверка сканирующего оборудования, подготовка вычислительной техники, формирование и организация каталогов, именование и запись полученных файлов на диск.

Хочется отметить, что работа по сканированию рукописного архива выполнялась в Фонограммархиве ИЯЛИ впервые. В настоящее время совместными усилиями фольклористов, архивистов и инженеров проводится работа по перепроверке, систематизации полученных файлов; намечаются сроки резервного копирования и проверки данных.