Уменьшение объема оцифрованных документов, размещаемых в Интернет

М.В. Дагаев

Российская государственная библиотека
costandy@list.ru

Аннотация

В последнее десятилетие в мире заметно выросли темпы оцифровки библиотечных и архивных фондов. Практически во всех странах запущены проекты перевода печатных изданий и рукописей в цифровую форму, рассчитанные на многие годы. Суммарное количество документов, запланированных к оцифровке, уже исчисляется десятками миллионов названий.

Но обратная сторона этого процесса в том, что так же быстро растут объемы данных, которые необходимо хранить в архивах проектов оцифровки и в электронных библиотеках. Даже в пределах одного крупномасштабного проекта объемы изображений считаются сотнями терабайт данных, а подобных проектов сейчас можно насчитать более десятка. В этой связи для электронных библиотек представляют постоянный интерес способы, которыми можно сократить размеры выкладываемых материалов.

В предлагаемой статье анализируются различные способы, позволяющие уменьшить объемы оцифрованных изображений.

То, что обычно называют оцифровкой — перевод печатных и рукописных документов в электронную форму — родилось на свет задолго до появления сканеров. Родоначальником оцифровки принято считать проект «Гутенберг» (США, университет штата Иллинойс, 1971) [1], но уже в 1960-х годах по советским ВЦ во множестве циркулировали магнитные ленты с «самиздатовскими» текстами.

Однако тот размах и всеохватность, которые мы наблюдаем вокруг себя сейчас, начался значительно позже — в 1980-х годах, когда появились на свет первые сканеры. К этому же времени относятся первые проекты массовой оцифровки. Один из наиболее известных примеров — «Archivo General de Indias», в котором хранятся документы, связанные с испанским завоеванием американского континента. Начиная с 1986 и по 1992 год, в связи с 500-летием открытия Америки, в нем было оцифровано около 8 млн. документов (цифра для технологических возможностей 1980-х годов совершенно фантастическая) [2].

В 1990-х годах у производителей сканирующего оборудования происходит что-то вроде тихой революции — появляется техника, ориентированная на быстрое, поточное, сканирование больших объемов печатных изданий и документов. Новые разновидности сканеров придают заметно б*о*льший размах и масштабность работам по оцифровке. Ближе к концу 1990-х, началу 2000-х годов в разных странах появляются крупномасштабные проекты с государственной поддержкой, рассчитанные на сроки в многие годы. В рамках таких проектов объемы, планируемые для обработки, считаются уже не миллионами страниц, а миллионами документов.

В Европе примером подобного может служить проект EDL, European Digital Library. Он был запущен в 2006 году, как система координации работ, проводимых в библиотеках и архивах стран ЕС. Фонд электронных документов EDL к данному моменту должен насчитывать более десяти миллионов отдельных документов.

В Японии хорошо известен проект Japan Center for Asian Historical Records, по переводу в оцифрованное состояние архивов эпохи Мэйдзи и более ранних времен. В 2002 г. в нем было более 2.5 млн. документов, к 2009 — более 6 млн. документов [7].

Среди работ по оцифровке, ведущихся в Китае, чаще всего упоминается China-US Million Book Digital Library Project. Изначально проект был рассчитан на обработку одного миллиона изданий., но давно уже перешагнул эту границу — на 2005 г. в его фондах было 1,1 млн книг, 450 млн страниц рукописных документов, 12 тыс. названий журналов, 600 названий газет.

В 2004 году заработал известный практически всем Google Book Search (тогда Google Print). На 2009 год в рамках этого проекта планировалось оцифровать около 30 млн. отдельных изданий.

Таким образом количество уже отсканированных документов исчисляется миллионами названий. Количество документов, запланированных для сканирования в ближайшее десятилетие, исчисляется еще большими цифрами — десятками миллионов названий.

В таблице 1 приведены примерные размеры изображений для некоторых, широко используемых форматов страниц. Ориентировочная оценка того,

Труды XIV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2011), Санкт-Петербург, Россия, 2011.

Таблица 1. Объемы данных, занимаемые изображениями различного размера (отсканированы в цвете, с разрешением в 300 dpi)

Геометрический размер страницы	Примеры изданий	Объем, в Мпикселах	Объем, в Мб
Формат A4 (210x297 мм)	Разворот книги, страница журнала	~8.7	~26.1
Формат A3 (297x420 мм)	Энциклопедия, художественный альбом	~17.4	~52.2
Формат A2 (420x594 мм)	Разворот газеты, картографический атлас	~34.8	~104.4
Формат A1 (594x840 мм)	Географическая карта, плакат	~69.6	~208.8
Формат A0 (840x1188 мм)	Географическая карта, плакат	~139.2	~417.6

сколько места будут занимать архивы изображений перечисленных проектов оцифровки, дает величины порядка сотен терабайт данных. Аналогичные цифры для содержимого электронных библиотек (ЭБ), где документы размещаются в сжатом виде, будут насчитывать десятки терабайт данных.

В подобной ситуации возникает вполне естественная потребность сократить размеры изображений, размещаемых в ЭБ, до возможного минимума. То есть уменьшить их настолько, насколько это позволяют сделать сегодняшние технологии, сохранив при этом необходимый уровень читаемости. Но, каким образом можно добиться подобного?

Возможны следующие варианты действий.

1. Более продуманное планирование того, в каком виде желательно сканировать и размещать в ЭБ различные типы изданий.

В настоящее время практика оцифровки нередко носит стихийный характер в части того, с каким разрешением, глубиной цвета и качеством надо сканировать и размещать в Интернет те или иные разновидности изданий. В результате документ чисто текстового содержания (свод законов) может быть отсканирован и выложен в ЭБ с тем же качеством изображения, которое требуются для художественного издания (альбома с репродукциями картин). Если при планировании процесса оцифровки учитывать то, для каких целей будет использоваться та или иная группа изданий, можно избежать неоправданного перерасхода места для их размещения.

2. Использование для выкладки более широкого набора графических форматов, чем это обычно практикуется.

Основным рабочим форматом для размещаемых в ЭБ изданий и документов считается PDF. Это формат т.н. «контейнерного» типа, в котором удобно размещать данные различных видов, в том числе и изображения. В настоящее время в нем выкладывается более 90% оцифрованных материалов [8].

Если выкладываемые изображения отсканированы в цвете или в Grayscale (полутонах серого), то перед сборкой в PDF их обычно переводят в JPEG, а если в Black&White (черно-белые) — то в формат TIFF G4. В то же время существуют и другие варианты преобразования в сжатый вид, которые способны обеспечить более высокие степени сжатия при сопоставимом качестве получаемых изображений.

3. Уменьшение объема растра, заключенного в изображении.

Материалы, проходящие через оцифровку, обычно сканируются с разрешением 300 dpi и более. Объем получаемых при этом изображений будет колебаться от 8-10 до 80-90 мегапикселов, в зависимости от геометрических размеров издания (см. таблицу 1). Но для того, чтобы обеспечить нормальную экранную читаемость полученных сканов, подобные пиксельные размеры часто избыточны. Их можно уменьшать, не теряя при этом в качестве восприятия. Таким образом, регулируя объема растра, заключенного в изображении, можно заметно сократить его размер еще до преобразования в сжатый вид.

Планирование процесса оцифровки

То, каким образом используются оцифрованные документы, можно разделить на следующие категории.

1. Документ представляет интерес только с точки зрения содержащейся в нем информации (текст, таблицы, графики, диаграммы и т.д). Сюда можно отнести научно-техническую литературу, большую часть неиллюстрированных изданий, ноты и т.д.

Возможные варианты использования таких документов: экранный просмотр и распознавание (OCR).

Требования к качеству оцифровки — должна сохраниться общая читаемость.

2. В документе ценна и интересна не только информационная составляющая, но и художественная (иллюстрации, фотографии профессионального качества, художественно выполненная верстка). Сюда можно отнести большую часть газет и журналов, иллюстрированные книги, художественные альбомы, книги для детей, открытки и т.д.

Возможные варианты использования таких документов: экранный просмотр, распознавание, использование для художественного оформления издательских работ (или же для полного издания/переиздания).

Требования к качеству оцифровки — внешний вид отсканированных страниц должен быть малоотличим от оригинала.

3. Кроме информационной и художественной составляющих, документ представляет интерес с исторической и/или научной точки зрения. Сюда можно отнести редкие и уникальные издания, старинные карты и атласы, рукописи и рукописные книги и т.д.

Возможные варианты использования таких документов: экранный просмотр, использование для художественного оформления издательских работ (или же для полного издания/переиздания), использование в качестве опорного материала при реставрационных работах, для научно-исследовательских работ.

Требования к качеству оцифровки — внешний вид отсканированных страниц должен полностью совпадать с оригиналом.

Перечисленные категории, если их оценивать с точки зрения объемов документов, образуют явно выраженную пирамиду.

Основная масса сканируемых материалов относится к первой категории и составляет массивное основание пирамиды. Поскольку здесь важна и существенна только общая читаемость страниц, то результаты сканирования вполне допустимо хранить в виде Black&White-изображений. Таким образом можно существенно уменьшить объемы данных, которые необходимо хранить в архивах и размещать в ЭБ. Получаемая экономия места может достигать 100 и более раз сравнительно с цветными изображениями, в зависимости от выбранного графического формата.

Гораздо меньшее, но тоже заметное количество сканируемых материалов составляет вторая категория — средняя часть пирамиды. Поскольку многие книги и журналы прежних лет издания не содержат цветных иллюстраций, то в этом случае резервом для сокращения объемов может стать глубина цвета. Если с точки зрения будущего использования таких материалов нет необходимости сохранять точный внешний вид страниц (т.е. хранить их в виде цветных изображений), то допустимо хранить их в виде Grayscale-изображений. Выигрыш в объемах данных получается не столь существенным, как в предыдущем случае — всего в три раза — но при больших объемах оцифровки он тоже может дать ощутимую экономию.

Материалы третьей категории требуется преобразовывать в оцифрованный вид с максимальной тщательностью и качеством. Здесь не представляется возможным выиграть что-либо на изменении политики оцифровки. С другой стороны издания и документы этого типа достаточно малочисленны и образуют самую верхнюю часть пирамиды.

Подборка изображений для тестирования

Чтобы проанализировать возможности различных форматов сжатия и то, что может дать регулирование объема растра, была создана тестовая подборка отсканированных материалов. В нее были подобраны характерные образцы тех изданий и документов, с которыми можно часто столкнуться при оцифровке фондов библиотек и архивов.

Содержимое подборки описано в таблицах 2 и 3.

Графические форматы и алгоритмы сжатия, которые можно использовать при работах по оцифровке

Black&White-изображения можно перевести в сжатый вид тремя основными способами — с помощью формата TIFF G4 (наиболее часто используемый вариант), с помощью алгоритма JBIG2 и с помощью формата DjVu. В таблице 4 приведены данные по сжатию нескольких книг с разным уровнем полиграфического качества.

Как можно видеть из этой таблицы, наилучшее сжатие Black&White-изображений дает формат DjVu. Он уверенно выигрывает по степени сжатия и у TIFF G4, и у JBIG2. Поэтому если основная масса материалов, размещенных в ЭБ, оцифрована в Black&White, то наиболее эффективным форматом для выкладки можно считать DjVu.

Таблица 2. Материалы, отсканированные в цвете

Разновидность документов	Источник	Время издания	Размер	Разрешение сканирования, dpi
Текст без иллюстраций (19 век)	Сборник законодательных актов	Конец 19 века	Близок к A4	300
Текст без иллюстраций (20 век)	Современный технический журнал	Конец 20 века	Близок к A4	600
Текст с иллюстрациями (19 век)	Архитектурный альбом	Середина 19 века	Близок к A2	300
Текст с иллюстрациями (20 век)	Современный технический журнал	Конец 20 века	Близок к A4	600
Иллюстрация	Иллюстрированный справочник	Конец 19 века	Близок к A3	300
Рукописный текст	Рукопись на старославянском	Конец 18 века	Близок к A4	300
Географическая карта	Карта озера Байкал	Середина 19 века	Близок к A1	300

Таблица 3. Материалы, отсканированные в Grayscale

Разновидность документов	Источник	Время издания	Размер	Разрешение сканирования, dpi
Текст без иллюстраций (18 век)	Географический словарь	Конец 18 века	Близок к А2	200
Текст без иллюстраций (20 век)	Математический журнал	Конец 20 века	Близок к А5	600
Ноты	Нотный альбом	Конец 19 века	Близок к А4	300
Фотография	Фотоальбом видов Москвы	Конец 19 века	Близок к А4	300

Таблица 4. Сравнение сжатия Black&White-изображений для форматов DjVu, TIFF G4 и JBIG2

Книга	Размеры изображений (Мб)			
	Несжатый TIFF	TIFF G4[1]	JBIG2	DjVu
Технический справочник	1 428	49	28	18
Записки о Туркестане	825	45	24	13
Адресная книга	1 640	110	77	46
История Казахстана	363	19	10	4

Когда речь идет о документах, отсканированных в цвете и Grayscale, то альтернативой формату JPEG может стать его «коллега» JPEG 2000 или же формат LDF (LuraTech Document Format).

Последний формат известен не столь широко, как JPEG и JPEG 2000, потому что программное обеспечение для него выпускает только один производитель — создатель формата, компания LuraTech [9]. Однако по критерию «коэффициент сжатия / качество получаемых изображений» он обычно превосходит другие форматы сжатия графики.

Сравнительные результаты по каждому из форматов можно увидеть в таблицах 5 и 6. Приводимые цифры соответствуют максимальному уровню сжатия исходных изображений, при котором еще сохраняется нормальная читаемость текста и возможность просматривать иллюстрации.

Для Grayscale-изображений переход с JPEG на другие форматы может дать выигрыш в объемах хранимых данных до 5-6 раз. Возможности JPEG 2000 и LDF по увеличению степени сжатия в этом случае будут примерно одинаковы.

Для цветных изображений выигрыш не настолько велик — до 2-3 раз. При этом лучшие результаты будет давать LDF, в особенности если речь идет об изданиях без иллюстраций (или с иллюстрациями несложного вида). На мелкодеталированных изображениях (карты, штриховые рисунки) разница в степени сжатия между JPEG 2000 и LDF становится минимальной.

Уменьшение объема растра, заключенного в изображении

Для оценки возможностей этого подхода были использованы изображения, перечисленные в таблице 2. Результаты сравнительного анализа приведены в таблице 7.

Таблица 5. Сравнение сжатия цветных изображений для форматов JPEG, JPEG 2000 и LDF

Вид документов	Несжатый TIFF, Мб	Объем (Мб) / Коэффициент сжатия		
		JPEG	JPEG 2000	LDF
Текст без иллюстраций (19 век)	23	0.71/32	0.32/72	0.23/100
Текст без иллюстраций (20 век)	79	1.31/60	1.01/78	0.39/202
Текст с иллюстрациями (19 век)	105	2.02/52	1.50/70	0.96/109
Текст с иллюстрациями (20 век)	94	1.61/58	1.17/80	0.46/204
Иллюстрация	43	0.80/54	0.57/75	0.39/110
Рукопись	24	0.59/41	0.34/70	0.23/104
Карта	136	3.01/45	1.89/72	1.52/90

Таблица 6. Сравнение сжатия Grayscale-изображений для форматов JPEG, JPEG 2000 и LDF

Вид документов	Несжатый TIFF, Мб	Объем (Мб) / Коэффициент сжатия		
		JPEG[1]	JPEG 2000	LDF
Текст без иллюстраций (19 век)	20	1.45/14	0.28/72	0.28/72
Текст без иллюстраций (20 век)	15	0.46/33	0.19/78	0.15/100
Нотный лист	10	0.84/12	0.18/56	0.18/56
Страница альбома	6.2	0.47/13	0.08/71	0.22/28

Таблица 7. Сравнение результатов сжатия отсканированных документов в исходном виде и после уменьшения пиксельного объема

Вид документов	Размер изображений				
	Исходный, TIFF		Конечный, TIFF		Исх./Кон., JPG
	Пикселов	Мб	Пикселов	Мб	Мб
Текст без иллюстраций (19 век)	2400 x 3305	23	1200 x 1679	5.8	0.94/0.18
Текст без иллюстраций (20 век)	4410 x 6257	79	1200 x 1702	5.8	1.75/0.29
Текст с иллюстрациями (19 век)	5193 x 7090	105	2200 x 3003	18.9	2.89/0.57
Текст с иллюстрациями (20 век)	4888 x 6712	94	1200 x 1647	5.7	2.10/0.31
Иллюстрация	3283 x 4596	43	1200 x 1679	5.7	1.15/0.18
Рукопись	2296 x 3600	24	1200 x 1881	6.5	0.82/0.26
Карта	7840 x 6080	136	3800 x 2946	32	4.02/1.20

Первоначальные размеры изображений, в пикселах и мегабайтах, даются в колонке «Исходный, TIFF». Затем каждое из них постепенно уменьшалось до тех пор, пока в нем еще сохранялась нормальная читаемость текста и возможность просмотра иллюстраций. Полученный к этому моменту размер, фиксировался в колонке «Конечный, TIFF».

После этого исходные и конечные изображения переводились из формата TIFF в JPEG (настройка сжатия Q=050), а полученные при этом размеры фиксировались в колонке «Исх./Кон., JPG».

Как можно увидеть из приводимой таблицы, для большей части изображений вполне допустимо уменьшение объема растра до 10-12 раз. Это дает возможность сократить размер итогового JPEG-файла в 5-6 раз, при том, что читаемость текстовой части и внешний вид иллюстраций остаются без видимых изменений.

Наилучшим образом сокращение пиксельного объема переносят страницы с содержимым «текст» или «текст плюс иллюстрации несложного вида». Здесь для оригиналов в несжатом TIFF можно добиться уменьшения размеров файлов до 10-12 раз и более (в зависимости от первоначальных значений), а для изображений, сжатых в JPEG — до 5-6 раз.

Наихудшим образом сокращение пиксельного объема переносят карты и рисунки, выполненные в стиле «гравюра». В этом случае можно получить только 5-6 кратное уменьшение для несжатого TIFF и 4-5 кратное — для JPEG-файлов.

При анализе процесса «уменьшение объема плюс сжатие» можно проследить достаточно хорошо выраженную закономерность. Хуже всего уменьшение растрового объема переносят документы, у которых на страницах много мелких, плотно расположенных деталей — текст, набранный небольшими кеглями, участки с плотной штриховкой, рисунки в «гравюрном» стиле. Если же на странице много плавных переходов цвета (полутоновые рисунки) и участков с однородной цветовой заливкой (текст, набранный большими кеглями, диаграммы), то уменьшение растрового объема дает гораздо лучшие результаты.

Литература

[1] Антопольский, А.Б. Электронные библиотеки: основные принципы создания: научно-методическое пособие. / А.Б. Антопольский, Т.В. Майстрович. – М.: «ЛИБЕРЕЯ» 2006.

[2] Браккер, Н.В. Инициативы Европейской комиссии по цифровым библиотекам. / Н.В. Браккер, Л.А. Куйбышев // Интернет и современное общество — 2006: Труды IX Всероссийской объединенной конференции – СПб., 2006. С. 6-12.

[3] Сванепоэль, М. Проекты оцифровки: изучение состояния дел в мире / М. Сванепоэль // Научные и технические библиотеки. 2009. № 3. С. 94-112.

[4] Muta, S. The Japan Center for Asian Historical Records: toward a full-fledged digital archives / S. Muta, A. Kobayashi // Joho Kanri. 2002. Vol. 45, № 7. P. 477-483.

[5] Zhao, J. Technical Issues on the China-US Million Book Digital Library Project / J. Zhao, C. Huang // Lecture Notes in Computer Science. 2005. Vol. 3334. P. 75-87.

[6] Сэломон, Д.. Сжатие данных, изображений и звука. М.: Техносфера, 2004.

[7] Сайт проекта «Japan Center for Asian Historical Records» [Электронный ресурс] — Режим доступа: http://www.jacar.go.jp/english/siryo/siryo2.html (дата обращения: 12.07.2011).

[8] Bültmann, B. Digitised Content in the UK Research Library and Archives Sector / B. Bültmann, R. Hardy, A. Muir, C. Wictor [Электронный ресурс] — Режим доступа: http://www.jisc.ac.uk/uploaded_documents/JISC-Digi-in-UK-FULL-v1-final.pdf (дата обращения: 12.07.2011).

[9] Сайт компании LuraTech [Электронный ресурс] — Режим доступа: http://www.luratech.com/en/home.html (дата обращения: 12.07.2011).

Reduction of volume of the digitized documents placed in the Internet

M. V. Dagaev

Within last ten years rates of digitisation of libraries and archives grow in the world all time. In all countries there are projects on scanning of printing editions and the manuscripts, planned on many years forward. For numbering many are planned many ten millions documents.

As a result as the quantity of the data which are required to be stored in long-preservation archives of digitisation and electronic libraries quickly grows. Total volumes of the digitized images make for a long time already hundreds terabyte. In this connection ways with which it is possible to reduce volumes of the data placed in digital libraries are of interest.

In article the ways are analyzed, allowing to reduce volumes of the digitized images.

[1] В тех случаях, когда изображения сжимаются с помощью TIFF G4 и JBIG2 они размещаются в ЭБ не в виде файлов соответствующих форматов, а в виде файла PDF, с упакованными в нем изображениями.