ЭЛЕКТРОННЫЕ РЕСУРСЫ. ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ

УДК 026.06+025.7/.9
DOI 10/33186/1027-3689-2019-4-53-60

В. И. Рожнов
ГПНТБ России

Хранение электронных книг в библиотеках: сравнительный анализ различных графических форматов

В статье приведены основные положения, связанные с оцифровкой документов. Дана краткая характеристика единых требований к оцифровке изданий для Национальной электронной библиотеки. Проанализированы графические форматы файлов, с помощью которых можно сохранять оцифрованные копии (TIFF, JPEG, PNG). Отмечены достоинства и недостатки каждого формата. Подчёркнуто, что в связи с быстрым развитием технологий и появлением устройств «автоматизированный читатель» библиотеки приобретают ставшее доступным сканирующее оборудование высокого разрешения и увеличивают скорость оцифровки с использованием тяжёлых графических форматов в своих фондах. Это позволяет наращивать темпы заполнения хранилищ серверов. Современные компьютерные технологии и программы распознавания текста позволяют начать оцифровку фонда. Автор задаёт закономерный вопрос: как затратить меньше ресурсов на оцифровку и дальнейшее хранение без потери культурной и исторической ценности документов? Предложено модифицировать классификацию документов и использовать технологию распознавания текста с последующим сохранением в файл, позволяющую работать с текстом напрямую (TXT, DOC, PDF). Автор приходит к выводу: существует объективная необходимость разработать новую технологию, лишенную недостатков уже существующих технологий, но сохраняющую их преемственность.

Ключевые слова: графические форматы, электронная книга, форматы для оцифровки фондов библиотек, TIFF, JPEG, PNG, TXT, DOC, PDF.

UDC 026.06+025.7/.9
DOI 10/33186/1027-3689-2019-4-53-60

Vladimir Rozhnov
Russian National Public Library for Science and Technology, Moscow, Russia

Storing digital books in electronic libraries: Comparison analysis of various graphic format use

The author begins with discussing the key provisions of document digitization. He characterizes the unified requirements of the National Electronic Library. Further, he analyzes the graphic formats for digitized copies (TIFF, JPEG, PNG). The advantages and drawbacks of each of the formats are discussed. Due to the rapid technologies advance and emergence of reader devices, the libraries that purchase now affordable HD-scanners can increase digitization efficiency with heavy graphic formats and to increase the rate of filling the servers' storage. The modern text recognition technologies and programs enable to initiate collections digitization. Having analyzed the formats, the author expectedly asks the following question: how to use less resources for digitization and further storage without losing the documents cultural and historical value? To solve this problem, the author suggests to modify document classification and use text recognition technologies with further storing in files which would enable to use the text directly (TXT, DOC, PDF). The author concludes that there is the need for a new consistent technology deprived of the existing drawbacks.

Keywords: graphic formats, digital book, formats for library collection digitization, TIFF, JPEG, PNG, TXT, DOC, PDF.

In 2013, the Russian State Library issued the "Uniform Requirements for Digitization of Publications…". They contain a classification of documents in three categories, recommendations on the equipment, requirements on the state of the document, the processing of master copies and the creation of user copies. The place and method of storing digitized materials is not specified. Recommended Tagged Image File Format as a raster image storage format. TIFF was created mainly to be compatible with various image processing devices. The advantage of TIFF is that it can support the full range of image sizes, resolutions and color depths as well as use of various compression methods. The main disadvantage of TIFF is the large file size resulting from the use of: a) lossless compression methods, b) tags for transferring image data. The large file size limits the use of TIFF in the educational environment of the Internet and slows down overall performance. Joint Photographic Experts Group format, JPEG is the name of a compression algorithm developed by an independent JPEG group. It uses bit information stored in pixel files, but it does not use indexed color. The advantage of the JPEG format is its ability to significantly compress large image files up to 1/20 of the original

file size. The Portable Network Graphics file format was created in 1995. The creators of PNG decided not to patent the product, but also searched for a flexible file format that would allow gamma correction for cross-platform consistency. They achieved their goals, and since 1996 the HTML format used to represent these files gain popularity. A PNG file has a block architecture that allows a flexible description of each attribute. In 1993, the Portable Document Format appeared and enables set passwords, digital signatures, etc. In 2017 our library has scanned 452,189 pages. The size of a digitized page is on average 3.5 MB using the Tiff format and a resolution of 300 dpi, which is equivalent to 1.52 TB of hard disk space. A year earlier, 359,765 pages were digitized, which is equivalent to 1.38 TB of volume. The load on the server space increases by 1.4 TB per year.

Электронные книги предназначены для считывания информации с электронно-совместимого устройства, планшета, смартфона или персонального компьютера. Электронные книги хранятся в виде электронных файлов разных форматов, они малы по объёму и легко распространяются.

С появлением цифровых технологий, устройств «автоматизированный читатель», возможности перевода изображения в цифровой вид библиотеки задумались о дальнейшей оцифровке и сохранности своих фондов.

В 1971 г. Майкл Харт, студент Университета штата Иллинойс (США), ввёл в компьютер текст Декларации независимости США прописными буквами и отправил сообщение ARPA.net (прародитель интернета). Это была первая электронная книга.

Объёмы фондов печатной продукции в мире огромные, и их перепечатка требует больших денежных и временных затрат. Возможность сканирования печатных носителей и разработка форматов изображения позволили сделать первый шаг к созданию электронных библиотек.

В 2013 г. Российская государственная библиотека выпустила «Единые требования к оцифровке изданий, включаемых в НЭБ». Они содержат классификацию документов по трём категориям, рекомендации по составу оборудования для оцифровки; отмечено, какими должны быть состояние документа перед оцифровкой, обработка мастер-копий и создание пользовательских копий по категориям. Место и способ хранения оцифрованных материалов в требованиях не прописаны.

Прежде чем убедиться в правильности сохранения электронных документов, необходимо провести анализ форматов, в которые их переводят. Рассмотрим графические форматы передачи данных.

TIFF (англ. *Tagged Image File Format*) – формат хранения растровых изображений; разработанный *Microsoft* и *Aldus* в 1986 г. *TIFF* является товарным знаком, первоначально зарегистрированным в *Aldus*, который впоследствии слился с *Adobe Systems* (Сан-Хосе, Калифорния, США). Теперь *Adobe* контролирует авторские права на спецификации *TIFF*.

TIFF был создан главным образом разработчиками изображений входных и выходных устройств, таких как принтеры, мониторы и сканеры; его предназначение – быть совместимым с различными устройствами обработки изображений.

Слово *Tagged* в аббревиатуре *TIFF* говорит о сложной файловой структуре этого формата. Первоначальный заголовок данных файла сопровождается «блоками» данных – тегами, которые передают информацию изображения в программу, отображающую файл.

Фактические спецификации *TIFF* определяют более 70 различных типов тегов. Этот уровень сложности обеспечивает большую гибкость между программами, однако программы, которые интерпретируют изображения *TIFF*, должны содержать все различные данные для тегов. Хотя многие программы упрощают это, реализуя только определённые теги, пропуск некоторых из них теоретически может повлиять на качество изображения, а частные теги могут ограничить использование файлов *TIFF* для некоторых приложений.

Наибольшее преимущество *TIFF* заключается в том, что он может поддерживать полный диапазон размеров изображений, разрешений и глубины цвета. Ещё одно преимущество – использование различных методов сжатия. Сжатие без потерь позволяет *TIFF*-файлам поддерживать разрешение изображения без потери детализации.

TIFF 5.0, выпущенный в 1988 г., включал поддержку технологии сжатия *LZW* (англ. *Lempel-Ziv-Welch* – Алгоритм Лемпеля – Зива – Велча). Хотя метод *LZW* – один из самых популярных алгоритмов сжатия, его использование может быть ограничено из-за его собственных ограничений, как отмечено выше. Одна из полезных особенностей *TIFF*-файлов – это то, что каждый может содержать более одного изображения.

Основной недостаток *TIFF* – большой размер файла, что является результатом использования: а) методов сжатия без потерь, б) тегов для передачи данных изображения. Большой размер файла ограничивает использование *TIFF* в образовательной среде интернета и замедляет общую производительность. Новейшая спецификация – *TIFF 6.0* (выпущена в 1992 г.) – включает метод сжатия *JPEG* с потерями, что позволяет уменьшить размер файла. В этой версии возникли некоторые проблемы, связанные с совместимостью, поскольку образы, использующие новые схемы сжатия, не могут быть успешно декодированы старым программным обеспечением. Однако этот недостаток, как ожидается, будет исправлен в *TIFF 7.0*.

JPEG (англ. *Joint Photographic Experts Group*) – Объединённая группа экспертов по фотографии – создан в начале 1990-х гг., следующее поколение схем сжатия файлов изображений.

JPEG – это не формат файла, а название алгоритма сжатия, разработанного независимой группой *JPEG*. Собственно формат файла *JPEG* называется *JPEG File interchange format* (*JIFF*), он создан специально для хранения и передачи фотографических изображений.

JPEG использует битовую информацию, хранящуюся в файлах пикселей, однако он не использует индексированный цвет. 24-битная цветовая схема файлов *JPEG* отображает каждый пиксель на экране с 24 битами кодирования данных, что даёт больше цвета и контрастности.

Преимущество формата *JPEG* – его способность значительно сжимать большие файлы изображений, что позволяет ускорить загрузку в электронной среде. В отличие от сжатия *LZW*, которое является типом линейного, или файлового, сжатия, более сложные алгоритмы *JPEG* позволяют выполнять истинное сжатие изображения. Схема сжатия дискретного косинусного преобразования *JPEG* делит изображение на 8x8-пиксельные секции и сжимает каждый раздел отдельно в три этапа.

Алгоритм дискретного косинусного преобразования используется во многих общих стандартах изображения и видео, включая группу экспертов по кинофильмам. С *JPEG* алгоритм математически сравнивает каждый пиксель со смежными пикселями, позволяя потребителю отрегулировать уровень сжатия. Сжатие изображения может быть достигнуто в соотношении до $^{1}/_{20}$ от исходного размера файла. Такие сжатые изображения могут быть переданы значительно быстрее в электронной среде и требуют существенно меньше места. Основной формат *JPEG* не запатентован.

Ещё одна полезная функция *JPEG* – прогрессивный дисплей, хотя он доступен только в новых веб-браузерах (впервые реализован в *Netscape Navigator 3.0* и *Internet Explorer 4.0*) и требует больше памяти. Прогрессивный дисплей позволяет потребителю увидеть предварительную версию низкого качества изображения до того, как полное изображение будет загружено.

Основной недостаток *JPEG* – потеря данных при каждом сжатии, что может привести к деградации изображения. Кроме того, процесс декодирования требует больше времени. Ещё один недостаток – возможное искажение, вызываемое техникой сжатия. Это может повлиять на файлы изображений, состоящие только из нескольких цветов или с большими областями одного цвета, такие как фон радиологического изображения. Другое возможное искажение – явление Гиббса, которое можно увидеть на изображениях с высоким разрешением, а также изображения с высокой пропускной способностью (например, линейное искусство).

Относительный новичок – формат файла ***PNG*** (англ. *Portable Network Graphics*, созданный в 1995 г. в ответ на собственный переход *GIF* (англ. *Graphics Interchange Format*) для обмена изображениями.

Создатели *PNG* приняли решение не патентовать продукт, но при этом вели поиск гибкого формата файла, который позволял бы без сжатия данных делать гамма-коррекцию для кроссплатформенной согласованности в яркости и прозрачности переменных. Они достигли своих целей, и с 1996 г. формат *HTML*, используемый для представления этих файлов, показал большую универсальность и продолжает набирать популярность. Следовательно, *PNG* обладает потенциалом, чтобы стать форматом файла изображения в будущем.

Файл *PNG* имеет блоковую архитектуру, которая подразумевает гибкое описание каждого атрибута и даёт *PNG* наибольшее теоретическое преимущество по сравнению с другими форматами файлов изображений. Широкий спектр информации изображения, а также поддержка буквенно-цифровых данных могут быть размещены в том же файле *PNG*.

Файл *PNG* состоит из первоначальной «подписи» (первые восемь байт), идентифицирующей файл именно как *PNG*-изображение, за которым следует серия блоков данных, кодирующих информацию изображения в программу-декодер (например, веб-браузер). Информация для изображения хранится в группах (блоках), отмеченных четырьмя именами персонажей, каждый из которых заканчивается «циклическим» (*CRC – cyclical redundancy check*), который проверяет целостность данных. Начальная подпись и *CRC* вместе с контрольной суммой *Adler-32* (алгоритм, аналогичный *CRC*, для несжатых данных) обеспечивают три типа проверки целостности файла *PNG*.

Данные блока могут содержать ключевые слова и строки информации, именуемые метаданными (т.е. информацией о данных), которые интерпретируются некоторыми программами декодера, игнорируемыми другими. Один из этих фрагментов обозначается как буквенно-цифровая текстовая строка, так что информация об изображении (например, аннотации, данные пациента, участник изображения) может храниться в этой строке. Следовательно, данные об изображении могут храниться в том же файле, что и само изображение. Поскольку поисковые системы с поддержкой метаданных интернета могут обнаруживать текстовые строки, изображение можно найти быстрее на основе имени файла изображения или текстового материала в текстовой строке. Архитектура блока файла *PNG* даёт возможность гибкого описания и частной информации на усмотрение автора.

Все рассмотренные выше форматы являются растровыми и занимают много места на жёстких дисках серверов. Объёмы отсканированной продукции растут в геометрической прогрессии. Возрастает актуальность изменения технологии и правил сохранности оцифрованных документов. Возника-

ют вопросы: почему не *PNG*-формат, а *JPEG*; почему не простой текстовый документ (распознанный) с разметкой является пользовательской копией; зачем нам сохранять страницы, не представляющие никакой ценности, в виде изображения, когда можно уменьшить размер, распознав её (перевести в векторную графику)?

Классификацию документов необходимо модифицировать в более конкретные подразделы: добавить информацию о популярности читаемых изданий, чтобы создать специальный архив и дать возможность удалить пользовательские копии, освободив место на серверах хранения; при обработке оцифрованных документов следует подходить к каждой единице более требовательно, а именно распознавать ту часть документа, которая не представляет культурную и историческую ценность.

Ретроконверсия – построение конвейерного производства, в котором процесс оцифровки разбит на отдельные воспроизводимые операции, а каждый шаг контролируется в режиме реального времени. Такой подход позволяет обрабатывать большие однотипные массивы документов с высокой эффективностью и заданным качеством.

В ГПНТБ России в 2017 г. отсканировано 452 189 страниц. Размер оцифрованной страницы – в среднем 3,5 Мб при использовании формата TIFF и разрешения 300 DPI, что эквивалентно 1,52 ТБ объёма на жёстком диске. Годом ранее было оцифровано 359 765 страниц, что эквивалентно 1,38 ТБ объёма. Нагрузка на пространство серверов прирастает по 1,4 ТБ в год.

Отдел сканирования и микрофильмирования располагает тремя планетарными сканерами формата А3, двумя – формата А2, одним – формата А1–А0 с установленной цифровой камерой; тремя барабанными сканерами для оцифровки карточек и листового материала формата А4.

Работу по оцифровке проводили шесть специалистов, в том числе два – по обработке оцифрованного материала. Готовый продукт поступает в систему хранения данных, основанную на кластеризации.

На рынке программного обеспечения представлены разработки, которые позволяют оцифровывать печатный текст. В России ведущей компанией является *ABBYY*; она предоставляет возможность распознать печатный текст на изображении и перевести его в текст за довольно короткое время. Конечный файл в сотни раз меньше графического оригинала, что впоследствии позволит во столько же раз сократить потребление физической памяти на жёстких дисках серверов.

Текстовый файл – это файл, содержащий текстовые данные, т.е. последовательность символов (в основном печатных знаков, принадлежащих тому или иному набору символов). Сам формат имеет ряд недостатков (нет воз-

можности интегрировать рисунки, таблицы и др.), а вот его наследники такую возможность имеют.

В 1990-х гг. корпорация *Microsoft* начала использовать расширение «.doc» для своего текстового процессора *Microsoft Word*. Изначально этот формат был определён как расширение имени файла, используемое для файлов, представляющих текст с разметкой или без.

В 1993 г. появляется формат *PDF* (*Portable Document Format*) – межплатформенный открытый формат электронных документов, изначально разработанный фирмой *Adobe Systems* с использованием ряда возможностей языка *PostScript*. В первую очередь формат предназначен для представления полиграфической продукции в электронном виде. Эволюция этого формата дала возможность устанавливать на документы пароли, водяные знаки, цифровые подписи, добавлять различные скрипты с возможностью самоуничтожения копии и т.п., добавлять шифрование, что позволит обезопасить документ от несанкционированного копирования.

Представленный в этой статье анализ показывает объективную необходимость разработки новой технологии, лишённой недостатков существующих технологий, но сохраняющей их преемственность. Оцифрованные фонды библиотек быстро заполняют физические пространства серверов. Многие издания содержат изображения, таблицы, формулы, автографы, иные пометки и надписи, что делает документ уникальным. Суть проработки технологического процесса – максимально уменьшить размер конечной (пользовательской) копии электронного файла, сохранив максимальное качество, быстрое предоставление пользователю запрашиваемого документа с соблюдением всех правовых норм и, главное, – сократить размер архивной копии.

Vladimir Rozhnov, *Head, System Software Department, Russian National Public Library for Science and Technology;*

sobaka@gpntb.ru

17, 3rd Khoroshevskaya st., 123298 Moscow, Russia