ЭЛЕКТРОННЫЕ ИНФОРМАЦИОННЫЕ РЕСУРСЫ

УДК 025.2 – 028.27 + 002.1 – 028.27
https://doi.org/10.33186/1027-3689-2026-10-50-71

Инструменты и методология обработки документов для долговременного архивирования в библиотеках

Е. В. Бескаравайная

Библиотека по естественным наукам РАН,
Москва, Российская Федерация,
elenabesk@gmail.com

Аннотация. В статье анализируются современные подходы к обработке документов в библиотеках и архивных учреждениях с акцентом на обеспечение дальнейшей долгосрочной сохранности цифровых материалов.

Целью исследования была разработка методологии систематизации разнородных цифровых материалов, хранящихся в фондах. Рассмотрены два основных сценария: подготовка отсканированных растровых изображений (нераспознанных документов) и нормализация цифровых документов в форматах DOCX, ODT и др. Особое внимание уделено использованию стандарта PDF/A-2b как оптимального для архивного хранения текстовых и смешанных документов.

Практическая значимость работы заключается в формировании алгоритма действий, включающего три основных этапа: предварительную оптимизацию исходных файлов (очистка, обрезка, нормализация шрифтов и метаданных); конвертацию в формат, рекомендованный для длительного хранения; последующую проверку полученных документов на соответствие требованиям стандарта ISO 19005. Важным итогом исследования стал подбор инструментов с открытым исходным кодом, таких как GraphicsMagick, Ghostscript, LibreOffice, OCRmyPDF и veraPDF, позволяющих автоматизировать работу с большим массивом документов.

В работе освещены вопросы совместимости предлагаемых инструментов с отечественным программным обеспечением и требованиями российского законодательства. Автор подчеркивает, что использование открытого ПО является для отечественных библиотек экономически оправданным.

Исследования выполнены при финансовой поддержке государственного задания Библиотеки по естественным наукам РАН 124112100065-5 (FFFN-2024-0003).

Ключевые слова: долговременное хранение, цифровые архивы, PDF/A-2b, нормализация документов, открытое программное обеспечение, библиотечные фонды, оцифровка

***Для цитирования*:** Бескаравайная Е. В. Инструменты и методология обработки документов для долговременного архивирования в библиотеках // Научные и технические библиотеки. 2026. № 10. С. 50–71. https://doi.org/10.33186/1027-3689-2026-10-50-71

DIGITAL INFORMATION RESOURCES

UDC 025.2 – 028.27 + 002.1 – 028.27
https://doi.org/10.33186/1027-3689-2026-10-50-71

Tools and methods of document processing for long-term archiving in libraries

Elena V. Beskaravainaya

Library for Natural Sciences, Russian Academy of Sciences,
Moscow, Russian Federation,
elenabesk@gmail.com

Abstract. The author analyzes modern approaches to document processing in libraries and archives with the focus on further long-term preservation of digital materials.

The purpose of the study is to develop classification methodology for heterogeneous digital materials in the collections. Two key scenarios are examined: obtaining scan rasters (unrecognized documents), or normalization of digital documents in DOCX, ODT, etc., formats. The emphasis is made on PDF/A-2b standard as an optimum for archiving text and hybrid documents.

The practical value of the study is that the processing algorithm is generated; it comprises three major stages: preliminary optimization of native files (file clearance, chopping, script and metadata normalization); conversion to recommended format for long-term storage; further checkup for ISO 19005 compliance. The significant result is selection of open source instruments, e. g. GraphicsMagick, Ghostscript, LibreOffice, OCRmyPDF, and veraPDF that enable big array processing automation.

The author also discusses compatibility of the mentioned instruments with Russian software and law provisions. She emphasizes that open source software is economically feasible for the Russian libraries.

The study is completed through the funding under the Government Order to the Library for Natural Sciences of the Russian Academy of Sciences 124112100065-5 (FFFN-2024-0003).

Keywords: long-term storage, digital archives, PDF/A-2b, document normalization, open source software, library collection, digitization

***Cite*:** Beskaravainaya E. V. Tools and methods of document processing for long-term archiving in libraries // Scientific and technical libraries. 2026. No. 3, pp. 50–71. https://doi.org/10.33186/1027-3689-2026-10-50-71

С развитием информационных технологий основными задачами поддержки и сохранения библиотечного фонда становятся его цифровое преобразование и обеспечение возможности, при которой документы могут быть найдены, открыты, прочитаны и поняты пользователями в будущем. Фонды библиотек содержат видео- и аудиоархивы, фотографии, карты, табличные данные, материалы исследований и экспериментов, но большую часть контента составляют текстовые документы на аналоговых и цифровых носителях, именно их перевод в форматы долговременного хранения является наиболее актуальным и нуждается в автоматизации. Такая работа требует, во-первых, четких подходов к архивированию документов, созданных в разное время и с помощью разного программного обеспечения, а во-вторых, привлечения программных продуктов, способных обеспечить автоматизацию

этих процессов в аспекте библиотечной работы. Если такие флагманы, как РГБ, РНБ, Президентская библиотека им. Б. Н. Ельцина, БАН, ГПНТБ, имеют возможность приобрести коммерческое ПО или создать собственное, то для региональных и муниципальных библиотек обработка цифрового контента – задача сложная, требующая планирования, ресурсов и специализированных знаний.

Целью нашего исследования была разработка практических алгоритмов действий при организации долговременного хранения цифровых материалов с применением доступного программного обеспечения.

Задачами исследования были:

– установить форматы текстовых документов на цифровых носителях, существующие в БЕН РАН;

– выявить форматы, перспективные для долгосрочного хранения в рекомендациях профессионального сообщества;

– разработать практические решения с перечнем доступного ПО для перевода документов в форматы, перспективные при организации долгосрочного хранения в библиотеках.

Ценность и практическая значимость исследования состоит прежде всего в составлении пошагового алгоритма для перевода имеющихся форматов в пригодные для длительного хранения, а также в определении набора инструментов, позволяющих автоматизировать процесс для больших объемов, начиная от анализа поступающего материала до тестирования результатов на соответствие требованиям долговременного хранения.

Критерии выбора форматов для долговременного хранения цифровых документов

Выбор подходящего формата – важнейшая стратегия в организации материалов на долгосрочное хранение. За последние 15 лет появилось множество зарубежных и отечественных исследований, направленных на оценку существующих форматов и возможности их участия в цифровом архивировании.

Ведущую роль в формировании рекомендаций играют отчеты, разработанные консорциумами и национальными учреждениями разных стран. Опыт их тестирований однозначно указывает на приоритет-

ность форматов, обеспечивающих баланс между качеством и совместимостью [**1**], позволяющих полностью восстановить исходный файл без потери качества или информации [**2**], способных к автоматическому контролю соответствия требованиям (валидируемости) [**3**]. Учреждения, занимающиеся организацией цифрового хранения на практике, принимают документы в едином цифровом формате (нормализация) [**4**] с наличием полного набора метаданных [**5**].

В правилах от 2023 г., разработанных Росархивом, основным и предпочтительным форматом для документов, предназначенных для долговременного хранения, указан PDF/A (п. 141), но рабочими версиями могут оставаться форматы DOCX или ODT, с возможностью последующей миграции в PDF/A [**6**].

Академическая литература рассматривает перспективные для хранения форматы на основе теоретических моделей. Еще в 2008 г. C. Becker [**6**] с коллегами, используя имитационное моделирование для прогнозирования «выживаемости» форматов в течение 30 лет, пришел к выводу, что открытые форматы имеют в 2–3 раза выше шансы остаться читаемыми, чем проприетарные. D. S. Rosenthal [**7**] и A. Brown [**8**] подчеркивали увеличение рисков устаревания, связанных не столько с форматом самим по себе, сколько с его поддержкой (наличие валидаторов, документации, сообщества разработчиков). D. Koepnick [**9**] показывал необходимость внедрения надежных конвейеров и разделение материала при архивировании на мастер- и производные копии. Е. В. Колесников с соавторами в процессе практического перевода архивных документов в цифровой формат сообщил о выборе в качестве рабочего TIFF-формата и сохранении многостраничных TIFF-файлов в виде базы на ленточных носителях [**11**]. А. В. Соловьев предложил собственную модель формата на основе языка XML, которая не только учитывает разнородность цифровых данных, но и позволяет организовать доступ к отдельным ее частям даже в случае отсутствия программных средств интерпретации при смене технологий [**12**].

Анализ литературы позволяет выделить универсальные критерии оценки при выборе формата для долговременного хранения:

1. Формат должен быть основан на открытом стандарте (ISO, OGC, IETF) и не зависеть от компании-разработчика или поставщика IT-продукции.

2. При переводе в цифровой формат должна быть обеспечена минимальная потеря информации (lossless-сжатие), либо исходные данные могут сохраняться в виде мастер-копии без деградации.

3. Важны валидируемость формата и наличие инструментов для автоматической проверки на соответствие выбранной спецификации.

4. Формат должен иметь возможность встраивать технические, административные и описательные метаданные.

Оценка форматов текстовых документов для долгосрочного архивирования

С развитием информационного общества состав фондов Библиотеки по естественным наукам РАН смещается с хранения информации на аналоговых носителях к цифровым (переведенным в цифровой формат) и электронным (созданным в цифровом формате). Материалы на цифровых носителях в библиотеке присутствуют в виде сканов (растровые изображения), документов с распознанным текстом и гибридных вариантов, совмещающих текст и изображения. Разнообразие форматов файлов и их версий приводит к необходимости поиска вариантов, которые были бы предпочтительны или приемлемы для долгосрочного хранения, в первую очередь это касается текстовых материалов в библиотеке.

На первом этапе работы мы определили список форматов документов, которые имеются в БЕН РАН, и на основе рекомендаций авторитетных учреждений, таких как Библиотека Конгресса США, Национальный архив Великобритании (PRONOM), Федеральное агентство по цифровым руководствам США (FADGI), Европейская комиссия по сохранению архивных документов и знаний (E-ARK), Росархив [**1**–**6**], составили описание достоинств и недостатков имеющихся форматов для длительного хранения (табл. 1).

Таблица 1

Описание форматов представления текстовых данных в перспективе долгосрочного хранения

Формат (расширение)	Достоинства для долговременного хранения	Недостатки для долговременного хранения
BMP (.bmp)	Растровый формат Microsoft; простая структура; отсутствие потерь качества (при хранении без сжатия); широкая программная совместимость	Огромный размер файлов, хранит данные пикселей без сжатия (неэффективное использование ресурсов); отсутствует поддержка метаданных; нет поддержки цветовых профилей ICC в базовой спецификации
TIFF (.tif, .tiff)	Стандарт для хранения растровых изображений; поддержка lossless-сжатия; богатые возможности для метаданных; поддержка ICC-профилей для управления цветом; поддержка многостраничности и цветовых пространств; рекомендуется стандартами	Сложность спецификации; может приводить к проблемам совместимости между программами; может включать проприетарные кодеки сжатия; большие размеры файлов при использовании сжатия без потерь
JPEG (.jpg, .jpeg)	Формат с высокой степенью сжатия; универсальная поддержка во всех системах; поддержка базовых метаданных	Необратимая потеря качества при каждом сохранении, критично для текста и графики; не подходит для глубокой цветокоррекции
PNG (.png)	Формат с lossless-сжатием без потерь качества; встроенная проверка целостности данных (CRC); открытая, документированная спецификация; удобен для графики, сканов с текстом, скриншотов	Большой размер файлов; ограниченная поддержка CMYK; слабая поддержка метаданных; не является предпочтительным форматом для полиграфии и архивирования
PSD (.psd)	Многослойный формат Adobe Photoshop; сохраняет слои, маски, позволяет вернуться к редактированию отдельных элементов; конвертирует в PDF	Проприетарный формат; зависимость от ПО; большие размеры файлов; не подходит для доступа и долгосрочного хранения; сложность валидации и миграции
DjVu (.djvu)	Формат для сканированных документов; очень хорошее сжатие для сканов; поддержка многослойности (текст + изображение); поддержка текстового слоя	Проприетарный; слабо поддерживается в современных ридерах; не стандартизован

Продолжение таблицы 1

Формат (расширение)	Достоинства для долговременного хранения	Недостатки для долговременного хранения
DOC (.doc)	Бинарный формат Microsoft; поддержка форматирования, интеграция с другими приложениями Word	Устаревший, высокий риск утраты данных; нет открытой спецификации; сложность валидации
DOCX / OOXML (.docx)	Форматированный текстовый открытый формат Microsoft; стандарт ISO/IEC; поддержка структуры и метаданных; может включать таблицы, стили, изображения; широкая совместимость; конвертируется в PDF	Сложная внутренняя структура; результат чтения в будущем зависит от библиотеки или программы (парсеров); риск потери форматирования при конвертации
EPUB	Формат электронных книг; открытый стандарт Консорциума Всемирной паутины (W3C); поддержка стилей, шрифтов, медиа, оглавления; возможность валидации; поддерживается большинством ридеров	Может включать средства защиты авторского права; сложная структура; не предназначен для научных данных или структурированного текста
HTML / XHTML (.html .xhtml)	Гипертекстовый формат с разметкой; открытый стандарт W3C; поддержка структуры стилей метаданных; легко конвертируется в PDF/A	Не самодостаточен (зависит от каскадной таблицы стилей CSS, JavaScript, внешних ресурсов); требует сохранения всей связки файлов (HTML, CSS, изображения); при архивировании веб-ресурсов переводится в контейнерный формат WARC (Web archive)
ODF (.odt .ods .odp)	Международный открытый стандарт для офисных документов ISO/IEC; самодостаточен (все данные внутри); поддержка структуры стилей метаданных; хорошая совместимость	Может терять форматирование при конвертации; для архива лучше конвертировать в PDF/A
PDF/A family	Семейство форматов для архивирования; стандартизован, встраивает шрифты и изображения, поддерживает метаданные, проверяется валидаторами	Может содержать javascript или внешние ссылки; требует нормализации; сложность при работе с интерактивными элементами

Окончание таблицы 1

Формат (расширение)	Достоинства для долговременного хранения	Недостатки для долговременного хранения
RTF (.rtf)	Полуоткрытый формат с простой разметкой; читается многими редакторами; поддерживает базовое форматирование	Устаревший; неполная поддержка Unicode; нестабильная реализация в разных программах; не самодостаточен
TXT (.txt)	Простой текст без форматирования; кодируется стандартами UTF-8, ASCII, CP1251 и др.; универсальный, читается всеми ПО; минимальный размер; никаких зависимостей; гарантирует доступность на любой системе в будущем	Нет структуры, заголовков, списков; нет метаданных; не подходит для документов с оформлением

Для каждого формата мы изучили рекомендации вышеупомянутых организаций и распределили их по балльной системе, где наивысшим является балл = 4, соответствующий «рекомендуемому формату», далее по уменьшению баллов: «рекомендуемый в отдельных случаях» = 3 балла; «допустимый» = 2 балла; «допустимый с ограничениями» = 1 балл. Как видно на рис. 1, в качестве основного предлагается формат PDF/A2, благодаря его совместимости с разным программным обеспечением, поддержке сложной графики, наличию надежных инструментов валидации, за ним следуют PDF/A-1 и XML. Хранение в ODF допустимо при невозможности использования PDF, а перевод и хранение на HTML/XHTML допускается только при наличии полного пакета ресурсов.

Что касается сохранения документов в форматах изображений, то рекомендованным всеми упомянутыми организациями остается TIFF без сжатия или с lossless-сжатием, PNG – допускается для определенных типов материалов, например, скриншотов и графиков, PSD – не рекомендуется как проприетарный, а JPEG и BMP – допускаются с ограничениями только для материалов с временным сроком хранения и не представляющих исторической ценности.

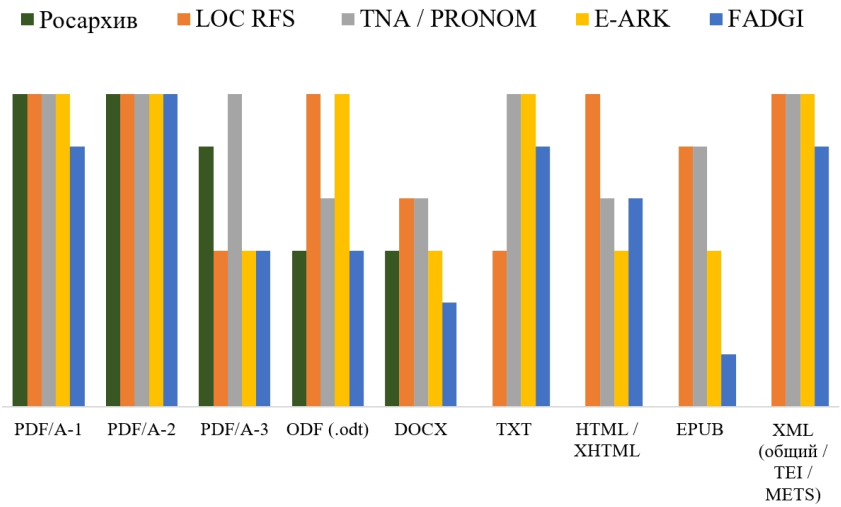


Рис. 1. Рекомендации ведущих организаций в области цифрового архивирования по выбору формата для длительного хранения

Вследствие того, что PDF/A-1 поддерживает только встроенные шрифты и растровые изображения, а для PDF/A-3 высок риск зависимости от проприетарных вложений, важнейшим вопросом становится миграция документов в формат PDF/A-2, способный сохранять внешний вид документа, независимо от инструментов создания, хранения и отображения файлов. Иногда в литературе встречаются дополнительные обозначения PDF/A-2a, PDF/A-2u, PDF/A-2b, отражающие уровни соответствия требованиям долгосрочного хранения, где (b) – базовый уровень, (u) – требует наличия карты Unicode для всех символов, (a) – включает перечисленные требования, плюс наличие дополнительных тегов, иерархии заголовков, описания изображений. Мы считаем, что для большинства задач сохранения в библиотеках вполне достаточно базового уровня, гарантирующего, что вид документа останется неизменным при воспроизведении в будущем.

Пошаговый алгоритм перевода текстовых документов в форматы для долгосрочного хранения

Процесс организации документов на длительное хранение включает три этапа:

1. Оптимизация исходного документа.
2. Конвертирование в рекомендуемый формат.
3. Проверка полученного документа на соответствие требованиям выбранной спецификации (верификация и валидация).

Материалы на цифровых носителях в библиотеках присутствуют в двух вариантах: оцифрованные копии в виде сканов и документы с распознанным текстом, отсюда различие в их обработке на начальном этапе.

Нераспознанные сканы (растровые изображения без текстового слоя) являются результатом масштабной работы библиотек по оцифровке документов собственного фонда. Поэтому еще на этапе сканирования необходимо:

– выбрать правильный цветовой режим (черно-белый для текстовых страниц, оттенки серого – для старой бумаги с рукописными пометками, цветной – для документов с иллюстрациями, штампами, диаграммами);

– выполнить обрезку для уменьшения размера файла;

– удалить белые поля, оставив 5–10 мм до текста, удалить развороты, переплет;

– исправить наклон страниц (выравнивание);

– удалить царапины, минимизировать цифровой шум.

Для сканирования многие российские библиотеки используют сегодня отечественное профессиональные планетарные и документные сканеры и ПО от «ЭларСкан». Среди сторонних программ для работы со сканерами стоит выделить открытое программное обеспечение с поддержкой пакетного режима: SANE (Scanner Access Now Easy), ScanTailor Advanced, gscan2pdf.

Автоматически обработать отсканированные нераспознанные документы, стандартизировать качество, размер, цвет и структуру изображений помогут инструменты, представленные в табл. 2.

Таблица 2

Инструменты для оптимизации растровых изображений

Инструмент	Что делает	Формат на выходе
ImageMagick (https://imagemagick.org/)	Анализ изображения, обрезка, поворот, управление цветностью	TIFF, PNG, JPEG, JPEG 2000, PDF
ScanTailor Advanced (https://scantailor.org/)	Разделение страниц, выравнивание угла, обрезка полей, улучшение контраста	TIFF, PNG, JPEG, WebP (в новых версиях)
GraphicsMagick (http://www.graphicsmagick.org/)	Обработка больших объемов изображений, поддержка многопоточности, автоматическая стандартизация качества, размера, цвета	TIFF, JPEG, PNG, GIF, BMP, PDF, SVG, PNM, WebP, HEIC, ограниченный JPEG 2000
Pillow (https://python-pillow.org/?spm=a2ty_o01.29997173.0.0.36145171Ziz5KJ)	Поворот, обрезка, нормализация яркости	JPEG, PNG, GIF, BMP, TIFF, WebP, ICO

Последним шагом оптимизации сканов являются удаление конфиденциальной информации и добавление базовых метаданных: технических (разрешение по горизонтали/вертикали, цветовой режим, метод сжатия, дата и время создания файла, инструмент обработки) и описательных (название документа, архивирующая организация, статус авторских прав). В табл. 3 мы собрали инструменты, которые помогут библиотекам провести работу с метаданными на массивах сканов в автоматическом режиме.

Сложной задачей для библиотек остается поиск инструментов для оптического распознавания символов (OCR-технологии), применимых к любому типу документов, независимо от языка, контекста и структуры. На наш взгляд, лучшим ПО продолжает быть коммерческий ABBYY FineReader как наиболее точный для распознавания кириллицы, сегодня он официально продается под названием ContentReader PDF и включен в Единый реестр российских программ под номером 17019. Среди открытых инструментов хорошо зарекомендовали себя: утилита OCRmyPDF от Tesseract (https://tesseract.projectnaptha.com/), способная накладывать текстовый слой поверх отсканированного PDF, сохраняя оригинальный вид страницы; система Kraken (https://kraken.re/main/

index.html), имеющая OCR-движок на платформах Linux и macOS для работы с редкими шрифтами и рукописями; пакет инструментов PaddleOCR (https://www.paddleocr.ai/main/en/index.html) для распознавания в сложных верстках и многоязычных документах. Все перечисленные инструменты полностью открыты и могут работать локально без передачи данных в облако, что критично в работе с персональными данными в библиотеках.

Таблица 3

Инструменты для работы с метаданными в сканах

Инструмент	Что делает
ExifTool	Чтение, запись, редактирование, удаление конфиденциальных метаданных в форматах TIFF, JPEG, PNG, PDF и др.
GIMP	Просмотр и ручное редактирование некоторых метаданных через GUI
MAT2 (Metadata Anonymisation Toolkit)	Анализ и удаление приватных метаданных, редактирование – специализированное
Droid/Siegfried	Извлечения базовых метаданных, редактирования нет
OCRmyPDF	Добавление/модификация метаданных
Archivematica	Автоматическое извлечение метаданных, редактирование

Для ценных и уникальных текстов очень важно перед конвертацией в PDF/A-2 сохранить сканы без сжатия в форматах TIFF, LZW (Lempel-Ziv-Welch) или JPEG 2000 (lossless). Во-первых, внесение исправлений после формирования PDF потребует полной пересборки документа, а во-вторых, в этот момент для таких материалов генерируется мастер-копия, хранящаяся отдельно от PDF. Несмотря на то что в отечественных официальных документах используются понятия «архивная копия» [**13**] и «электронный образ» [**14**], принципы их создания, параметры и хранение в точности соответствуют таковым для технического термина «мастер-копия», который в контексте архивирования обозначает эталонный экземпляр, обладающий максимальной полнотой информации и качеством [**15**]. На первый взгляд, сохранение двух копий одного документа требует дополнительных затрат, тем не менее оно обеспечивает экономию ресурсов в будущем: 1) мастер-копия

полностью соответствуют оригиналу и позволяет восстановить документ при утрате носителя; 2) она хранится в открытом формате, поэтому имеет больше шансов быть прочитанной при смене технологии; 3) по подсчетам французских коллег [**16**], процесс повторной оцифровки стоит в 3–5 раз дороже, чем хранение мастер-копии. Следует упомянуть о том, что библиотека не только хранит информацию, но и организует доступ к ней, поэтому, в отличие от мастер-копии, служащей для долговременного сохранения полной информации, необходима оптимизированная версия документа, совместимая, например, с мобильными технологиями или адаптированная для людей с ограниченными возможностями.

Рассмотрим второй вариант текстовых документов в фонде библиотеки – файлы ODT, DOCX, TXT. Такие форматы имеют распознанный текстовый слой, и перед конвертацией для них потребуется стандартизация оформления:

удаление ненужных вложений, макросов, гиперссылок, плавающих объектов, комментариев;

проверка авторских прав;

проверка корректности метаданных; проверка соответствия шрифтов внедрению в PDF.

Отдельного внимания потребует наличие внутри текстовых документов изображений, диаграмм, Excel-таблиц, формул. Когда процесс сканирования в библиотеке только начинался, очень перспективным казался DjVu, успешно работающий с гибридным текстом, сочетающий хорошую читаемость, возможность поиска по тексту, достаточную визуализацию изображения, при том что размер файла был в 5–10 раз меньше, чем у эквивалентных JPEG или PDF. Однако «формат будущего» не прижился: закрытое ПО ограничило эволюцию инструментов, а развитие технологий сжатия изображений и распознавания текста в PDF привело к тому, что международные организации сегодня не включают DjVu в список рекомендуемых форматов для сохранения, хотя он все еще используется для хранения старых коллекций и для чтения книг.

Для того чтобы результат нашей работы соответствовал принципам долговременного хранения, обозначенных в российских и международных стандартах, перед конвертацией рисунки и диаграммы

должны иметь разрешение не менее 300 dpi, встроенные объекты заменены на статичные изображения, формулы должны быть пересозданы с помощью стандартных редакторов формул с возможностью текстового поиска и копирования.

После оптимизации разных видов документов мы подошли к этапу **конвертирования** в рекомендуемый для длительного хранения формат PDF/A-2. Пакетную конвертацию больших коллекций в фоновом режиме помогут осуществить, например, LibreOffice или Ghostscrip (https://ghostscript.com) – бесплатные кроссплатформенные инструменты с открытым исходным кодом. Они проверят на соответствие шрифты, вырежут объекты JavaScript, внешние ссылки и мультимедиа, обработают метаданные.

Выявить ошибки конвертации помогут Apache Preflight (https://pdfbox.apache.org/) или veraPDF (https://docs.verapdf.org). Они проверят соблюдение технических требований стандарта ISO 19005 (**верификация**) для всех форматов PDF/A, включая WTPDF (Web Technology PDF); оценят структуру, корректность текста, полноту, читаемость (**валидация**).

Поэтапный алгоритм работы с текстовыми документами в разных цифровых форматах с применением открытых инструментов, способных автоматически обрабатывать большие массивы, представлен на рис. 2.

Следует подчеркнуть, что мы не просто переводим документы в формат PDF/A-2, мы направляем их на длительное хранение, отсюда повышенное внимание к чистоте метаданных и спецификации. Согласно международному стандарту метаданных PREMIS (Preservation Metadata: Implementation Strategies), из цифровых объектов должны быть удалены персональные данные, служебная информация, скрытый контент, сведения о геолокации, а вся история изменений файла зафиксирована в системе [**17**].

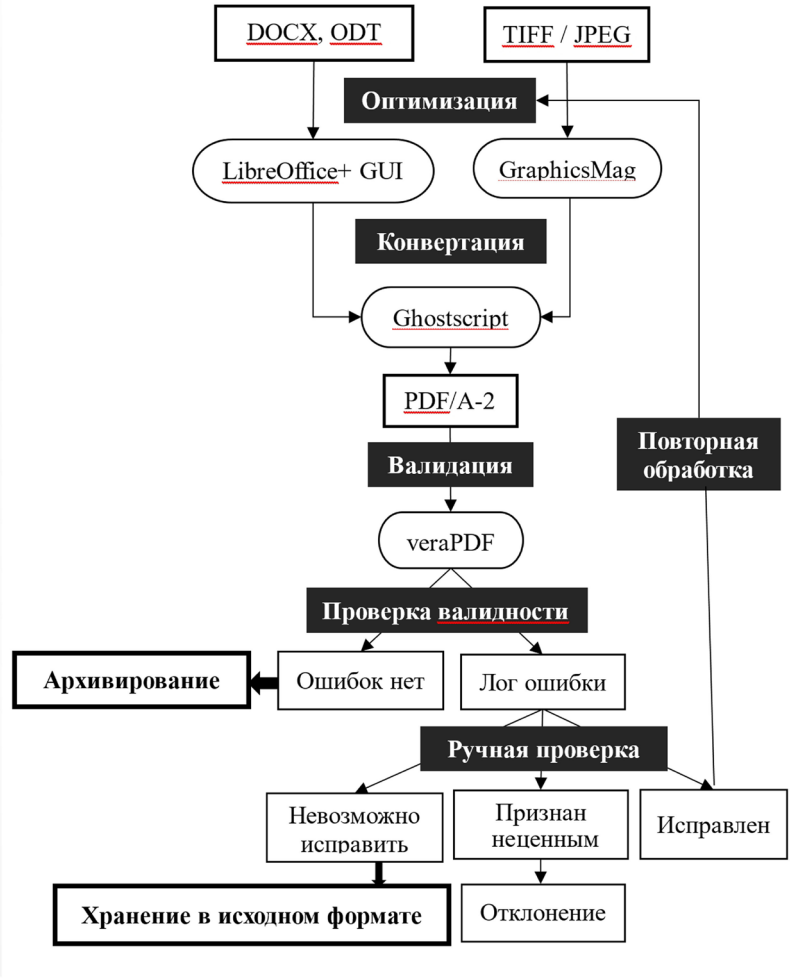


Рис. 2. Поэтапный алгоритм обработки документов при организации длительного хранения

Разработка отечественных инструментов для генерации и конвертации документов идет по двум направлениям. Во-первых, создание

собственных ПО, таких как: редакторы документов «Автономный модуль редактирования» в «МойОфис» и Document Builder в «Р7-Офис»; модули от системы ЭЛАР – «Саперио», который валидирует PDF/A по аналогии с veraPDF, и «Контекст», обрабатывающий метаданные подобно ExifTool.

Во-вторых, внедрение сторонних открытых инструментов на сертифицированные российские ОС (Astra Linux SE/CE, RedOS, РОСА), при котором они становятся частью среды и их использование правомерно. Среди них стандартные компоненты Linux, схожие с Ghostscript и GraphicsMagick; модули «Формат» для обработки графики и «Конвертор» для извлечения метаданных в составе АИС «Платформа» Росархива; программные технологии, совместимые с LibreOffice, позволяющие обрабатывать документы и конвертировать в PDF/A-2 (записи в реестре российского ПО № 14404 и 21479).

Однако приобретение всех перечисленных выше готовых программных продуктов потребует существенных вложений, поэтому, на наш взгляд, при организации долгосрочного хранения в библиотеках перспективным остается внедрение таких инструментов, как veraPDF, Ghostscript, ExifTool, GraphicsMagick, LibreOffice, – они относятся к категории ПО с открытым исходным кодом (Open Source), используются без коммерческой поддержки со стороны иностранных лиц, не подпадают под запрет российского законодательства [**18**], а если мы не собираемся встраивать их код в платные приложения, то их лицензии позволяют пользоваться инструментами бесплатно.

Выводы

1. Процесс организации длительного хранения начинается на этапе отбора документов в соответствии с критериями историко-культурной ценности, научной или социальной значимости. Для отечественных библиотек в условиях ограниченных ресурсов разумным становится подход, при котором максимальное внимание отдается сохранению уникальных документов в устаревших форматах, которые должны быть оцифрованы в высоком разрешении с умеренным сжатием, сохранены в форматы, доступные для миграции, проверены на отсутствие потерь и ошибок вручную, а при хранении иметь как мастер-копию, так и пользовательский вариант.

2. Для основного массива цифровых документов организация длительного хранения должна строиться на принципе «разумной целесообразности», при котором применяется автоматизированная обработка и выборочный контроль. Для экономии ресурсов допустимо оставлять на хранение сканы отдельных документов в формате JPEG 2000 как необходимом и достаточном, например для газет и карт [**19**, **20**].

3. Этап оптимизации перед конвертацией является критическим. Здесь стоит обратить внимание на неявный, но значимый фактор: попытка «сэкономить» на этапе оптимизации может привести к созданию технических PDF, реально непригодных для долгосрочного хранения. После конвертации любые дефекты исходного изображения нельзя устранить, шумы и тени не позволят правильно распознать текст, а неверный выбор цветового пространства приведет к будущим проблемам при просмотре на разных устройствах или при печати.

4. Российская нормативная база (Росархив) требует соответствия стандарту PDF/A [**10**] в целом, без привязки к конкретному уровню. На наш взгляд, для библиотечной практики стоит рассмотреть формат PDF/A-2b: он поддерживается современными технологиями, обладает хорошим сжатием, умеет объединять несколько файлов в один PDF-контейнер с общей обложкой и навигацией, что особенно актуально для архивирования многотомных изданий и последующего описания в MARC21 или RUSMARC.

5. Один из самых сложных этапов в организации длительного хранения на практике – выбор системы инструментов, способных автоматизировать многочасовую рутинную работу сотрудника, и интеграция в них. Лучшими, на наш взгляд, являются следующие открытые инструменты: GraphicsMagick – для обработки растровых изображений; пакет LibreOffice – для замены шрифтов, редактирования текстов и таблиц; набор программного обеспечения Ghostscript – для нормализации сконвертированных в PDF/A-1/2/3 файлов, а для их валидации на соответствие стандарту ISO 19005 – инструмент veraPDF.

Было бы слишком оптимистичным надеяться, что переход на открытые инструменты произойдет без привлечения IT-специалистов. Пример из практики: многие российские архивы в 2010-х годах закупили Adobe Acrobat Pro для создания архивных копий, в результате в локальных регламентах закрепилось требование: «…архивная копия

должна иметь признак создания в лицензионном ПО Adobe». Замена коммерческого Acrobat Pro на Ghostscript потребовала приглашения IT-инженера и дополнительной настройки системы, начиная от распознавания встроенных шрифтов из русскоязычных гарнитур до установки дополнительных утилит и скриптов для автоматической приемки в хранилище. В итоге реальная стоимость внедрения включала: приглашение программиста для первоначальной настройки, обучение персонала, повторное приглашение IT-специалиста для тестирования и анализа работы системы, доработки контроля. Тем не менее для российских библиотек переход на инструменты с открытым кодом и спецификацией – лучший, а иногда и единственно оправданный выход: закупки отечественных коммерческих продуктов – это дорого, а если ПО иностранное, проприетарное, покупка нарушает требования законодательства РФ.

Мы надеемся, что представленный подход с четкими рекомендациями и перечнем доступных открытых программных средств поможет библиотекам ускорить организацию длительного хранения цифровых фондов.

Список источников

1. **Technical** Guidelines for Digitizing Cultural Heritage Materials / Federal Agencies Digitization Guidelines Initiative (FADGI). Washington, DC : U.S. National Archives and Records Administration, 2020. URL: https://www.digitizationguidelines.gov/guidelines/ (дата обращения: 15.01.2026).

2. **E-ARK** Consortium. Recommended Practices for Preservation. Deliverables D4.2 – D4.4. Brussels : E-ARK Project, 2017–2020. URL: https://www.eark-project.com/resources/deliverables (дата обращения: 15.01.2026).

3. **Format** Risk Registry**:** оценка долгосрочной жизнеспособности цифровых форматов // Open Preservation Foundation : [сайт]. URL: https://openpreservation.org/resources/format-risk-registry/ (дата обращения: 03.03.2026).

4. **PRONOM**: The File Format Registry // The National Archives (UK) : [сайт]. URL: https://www.nationalarchives.gov.uk/PRONOM (дата обращения: 16.01.2026).

5. **Draft** National Archives Strategic Plan / U.S. National Archives and Records Administration. Washington, DC, 2017. URL: https://www.archives.gov/files/about/plans-reports/strategic-plan/strategic-plan-framework-2026-2030.pdf (дата обращения: 03.02.2026).

6. **Becker C., Kulovits H., Rauber A.** Evaluating Strategies for Digital Preservation: A Simulation-Based Approach // Proceedings of the 5th International Conference on Preservation of Digital Objects (IPRES 2008). Washington, DC, 2008. P. 112–121.

7. **Rosenthal D. S. H.** Format Obsolescence and Digital Preservation // IEEE Annals of the History of Computing. 2015. Vol. 37, no. 3. P. 40–50.

8. **Brown A.** Practical Digital Preservation: A How-To Guide for Organizations of Any Size. London : Facet Publishing, 2018. 272 p.

9. **Koepnick B.** Preserving Born-Digital Video: A Practical Guide. Washington, DC : CLIR, 2021. 124 p. (CLIR Report ; no. 186). URL: https://clir.org/pubs/reports/pub186 (дата обращения: 17.01.2026).

10. **Об утверждении** Правил организации хранения, комплектования, учета и использования документов Архивного фонда Российской Федерации... : Приказ Федерального архивного агентства от 31.07.2023 № 77 // Официальный интернет-портал правовой информации. URL: http://pravo.gov.ru (дата обращения: 06.09.2023).

11. **Колесников Е. В., Шаймарданов В. М.** Создание автоматизированной информационной технологии сканирования бумажных документов Росгидромета // Успехи кибернетики. 2022. Т. 3, № 1. С. 28–38. DOI 10.51790/2712-9942-2022-3-1-4.

12. **Соловьев А. В.** Решение проблемы интерпретации цифровых данных долговременного хранения // Труды Института системного анализа РАН. 2021. Т. 71, № 2. С. 43–49. DOI 10.14357/20790279210206.

13. **Об архивном** деле в Российской Федерации : федеральный закон от 22.10.2004 № 125-ФЗ : (с изм. от 24.07.2023) // Официальный интернет-портал правовой информации. URL: publication.pravo.gov.ru (дата обращения: 29.01.2026).

14. **Об утверждении** Требований к созданию, хранению и использованию электронных образов документов... : Приказ Министерства культуры РФ от 11.11.2022 № 746 // КонсультантПлюс : [справочно-правовая система]. URL: www.consultant.ru (дата обращения: 29.01.2026).

15. **Электронный** фонд пользования: создание, хранение, учет и использование : метод. рекомендации / ВНИИДАД ; исполн. Н. Г. Суровцева [и др.]. Москва, 2023.

16. **Bibliothèque** nationale de France. Plan stratégique 2020–2024 «Préserver et transmettre». Paris, 2020. URL: https://www.bnf.fr/sites/default/files/2020-11/plan_strategique_2020-2024.pdf (дата обращения: 29.01.2026).

17. **PREMIS** Data Dictionary for Preservation Metadata, Version 3.0 / PREMIS Editorial Committee. Washington, DC : Library of Congress, 2015. 274 p. URL: www.loc.gov (дата обращения: 10.02.2025).

18. **О мерах** по обеспечению технологической независимости и безопасности критической информационной инфраструктуры Российской Федерации : Указ Президента РФ от 30.03.2022 № 166 // Собр. законодательства Рос. Федерации. 2022. № 14. Ст. 2242.

19. **Technical** Guidelines for Digitisation at KB / KB, National Library of the Netherlands. The Hague, 2020. 23 p. URL: www.kb.nl (дата обращения: 23.01.2026).

20. **JPEG 2000** Bitstream File Format (ISO 15444-1) // Sustainability of Digital Formats: Planning for Library of Congress Collections / Library of Congress. Washington, DC, 2023. URL: www.loc.gov (дата обращения: 23.01.2026).

Reference

1. **Technical** Guidelines for Digitizing Cultural Heritage Materials / Federal Agencies Digitization Guidelines Initiative (FADGI). Washington, DC : U.S. National Archives and Records Administration, 2020. URL: https://www.digitizationguidelines.gov/guidelines/ (дата обращения: 15.01.2026).

2. **E-ARK** Consortium. Recommended Practices for Preservation. Deliverables D4.2 – D4.4. Brussels : E-ARK Project, 2017–2020. URL: https://www.eark-project.com/resources/deliverables (дата обращения: 15.01.2026).

3. **Format** Risk Registry**:** оценка долгосрочной жизнеспособности цифровых форматов // Open Preservation Foundation : [сайт]. URL: https://openpreservation.org/resources/format-risk-registry/ (дата обращения: 03.03.2026).

4. **PRONOM**: The File Format Registry // The National Archives (UK) : [сайт]. URL: https://www.nationalarchives.gov.uk/PRONOM (дата обращения: 16.01.2026).

5. **Draft** National Archives Strategic Plan / U.S. National Archives and Records Administration. Washington, DC, 2017. URL: https://www.archives.gov/files/about/plans-reports/strategic-plan/strategic-plan-framework-2026-2030.pdf (дата обращения: 03.02.2026).

6. **Becker C., Kulovits H., Rauber A.** Evaluating Strategies for Digital Preservation: A Simulation-Based Approach // Proceedings of the 5th International Conference on Preservation of Digital Objects (IPRES 2008). Washington, DC, 2008. P. 112–121.

7. **Rosenthal D. S. H.** Format Obsolescence and Digital Preservation // IEEE Annals of the History of Computing. 2015. Vol. 37, no. 3. P. 40–50.

8. **Brown A.** Practical Digital Preservation: A How-To Guide for Organizations of Any Size. London : Facet Publishing, 2018. 272 p.

9. **Koepnick B.** Preserving Born-Digital Video: A Practical Guide. Washington, DC : CLIR, 2021. 124 p. (CLIR Report ; no. 186). URL: https://clir.org/pubs/reports/pub186 (дата обращения: 17.01.2026).

10. **Ob utverzhdenii** Pravil organizatcii khraneniia, komplektovaniia, ucheta i ispol`zovaniia dokumentov Arhivnogo fonda Rossii`skoi` Federatcii... : Prikaz Federal`nogo arhivnogo agentstva ot 31.07.2023 № 77 // Ofitcial`ny`i` internet-portal pravovoi` informatcii. URL: http://pravo.gov.ru (data obrashcheniia: 06.09.2023).

11. **Kolesnikov E. V., Shai`mardanov V. M.** Sozdanie avtomatizirovannoi` informatcionnoi` tekhnologii skanirovaniia bumazhny`kh dokumentov Rosgidrometa // Uspehi kibernetiki. 2022. T. 3, № 1. S. 28–38. DOI 10.51790/2712-9942-2022-3-1-4.

12. **Solov`ev A. V.** Reshenie problemy` interpretatcii tcifrovy`kh danny`kh dolgovremennogo khraneniia // Trudy` Instituta sistemnogo analiza RAN. 2021. T. 71, № 2. **S**. 43–49. DOI 10.14357/20790279210206.

13. **Ob arhivnom** dele v Rossii`skoi` Federatcii : federal`ny`i` zakon ot 22.10.2004 № 125-FZ : (s izm. ot 24.07.2023) // Ofitcial`ny`i` internet-portal pravovoi` informatcii. URL: publication.pravo.gov.ru (data obrashcheniia: 29.01.2026).

14. **Ob utverzhdenii** Trebovanii` k sozdaniiu, khraneniiu i ispol`zovaniiu e`lektronny`kh obrazov dokumentov... : Prikaz Ministerstva kul`tury` RF ot 11.11.2022 № 746 // Konsul`tantPlius : [spravochno-pravovaia sistema]. URL: www.consultant.ru (data obrashcheniia: 29.01.2026).

15. **E`lektronny`i`** fond pol`zovaniia: sozdanie, khranenie, uchet i ispol`zovanie : metod. rekomendatcii / VNIIDAD ; ispoln. N. G. Surovtceva [i dr.]. Moskva**,** 2023.

16. **Bibliothèque** nationale de France. Plan stratégique 2020–2024 «Préserver et transmettre». Paris, 2020. URL: https://www.bnf.fr/sites/default/files/2020-11/plan_strategique_2020-2024.pdf (дата обращения: 29.01.2026).

17. **PREMIS** Data Dictionary for Preservation Metadata, Version 3.0 / PREMIS Editorial Committee. Washington, DC : Library of Congress, 2015. 274 p. URL: www.loc.gov (дата обращения: 10.02.2025).

18. **O merakh** po obespecheniiu tekhnologicheskoi` nezavisimosti i bezopasnosti kriticheskoi` informatcionnoi` infrastruktury` Rossii`skoi` Federatcii : Ukaz Prezidenta RF ot 30.03.2022 № 166 // Sobr. zakonodatel`stva Ros. Federatcii. 2022. № 14. St. 2242.

19. **Technical** Guidelines for Digitisation at KB / KB, National Library of the Netherlands. The Hague, 2020. 23 p. URL: www.kb.nl (дата обращения: 23.01.2026).

20. **JPEG 2000** Bitstream File Format (ISO 15444-1) // Sustainability of Digital Formats: Planning for Library of Congress Collections / Library of Congress. Washington, DC, 2023. URL: www.loc.gov (дата обращения: 23.01.2026).

Информация об авторе / Author

Бескаравайная Елена Вячеславовна – старший научный сотрудник Библиотеки по естественным наукам РАН, Москва, Российская Федерация

elenabesk@gmail.com

Elena V. Beskaravainaya – Senior Researcher, Library for Natural Sciences, Russian Academy of Sciences, Moscow, Russian Federation

elenabesk@gmail.com