empenoso
10 мая в 05:42

Как убрать пустые оборотные страницы из PDF после двухстороннего сканирования

Средний 6 мин 4.6K

Open source*, PDF, Софт, Лайфхаки для гиков

Кейс

Около двух месяцев назад я написал статью как сканировать многостраничные двухсторонние документы, если под рукой только обычный сканер с автоподачей, в которой затронул проблему того, что МФУ часто имеют дуплексную двустороннюю печать, но односторонний сканер.

Однако после решения проблемы быстрого сканирования больших двухсторонних документов, была обнаружена ещё одна проблема — некоторое количество страниц могут оказаться односторонними. И это означает, что PDF будет иметь белые страницы, например, со сканами перфораций или отверстий под кольца.

Конечно, можно удалить несколько страниц из PDF вручную, но что если таких файлов сотни, а сами документы имеют несколько десятков или даже сотен страниц как на фотографии?

Большой многостраничный документ

▸ TL;DR

Вариант удаления пустых страниц из pdf при помощи локальной программы

Перед тем как начать писать свой скрипт я честно пытался разобраться как удалить пустые страницы из пдф при помощи штатных средств какой-нибудь программы:

1. Пытался сделать это при помощи бесплатной открытой PDFsam Basic, которая доступна под Linux и Windows, и MacOS, потому что в интернете нашёл инструкции, но они оказались устаревшими.
2. Пытался сделать это при помощи Adobe Acrobat Pro, но у меня не получилось. Делал по инструкции:
    1. Откройте файл PDF в Adobe Acrobat.
    2. Нажмите на вкладку «Инструменты» в верхней строке меню.

2. Нажмите на вкладку «Инструменты» в верхней строке меню.

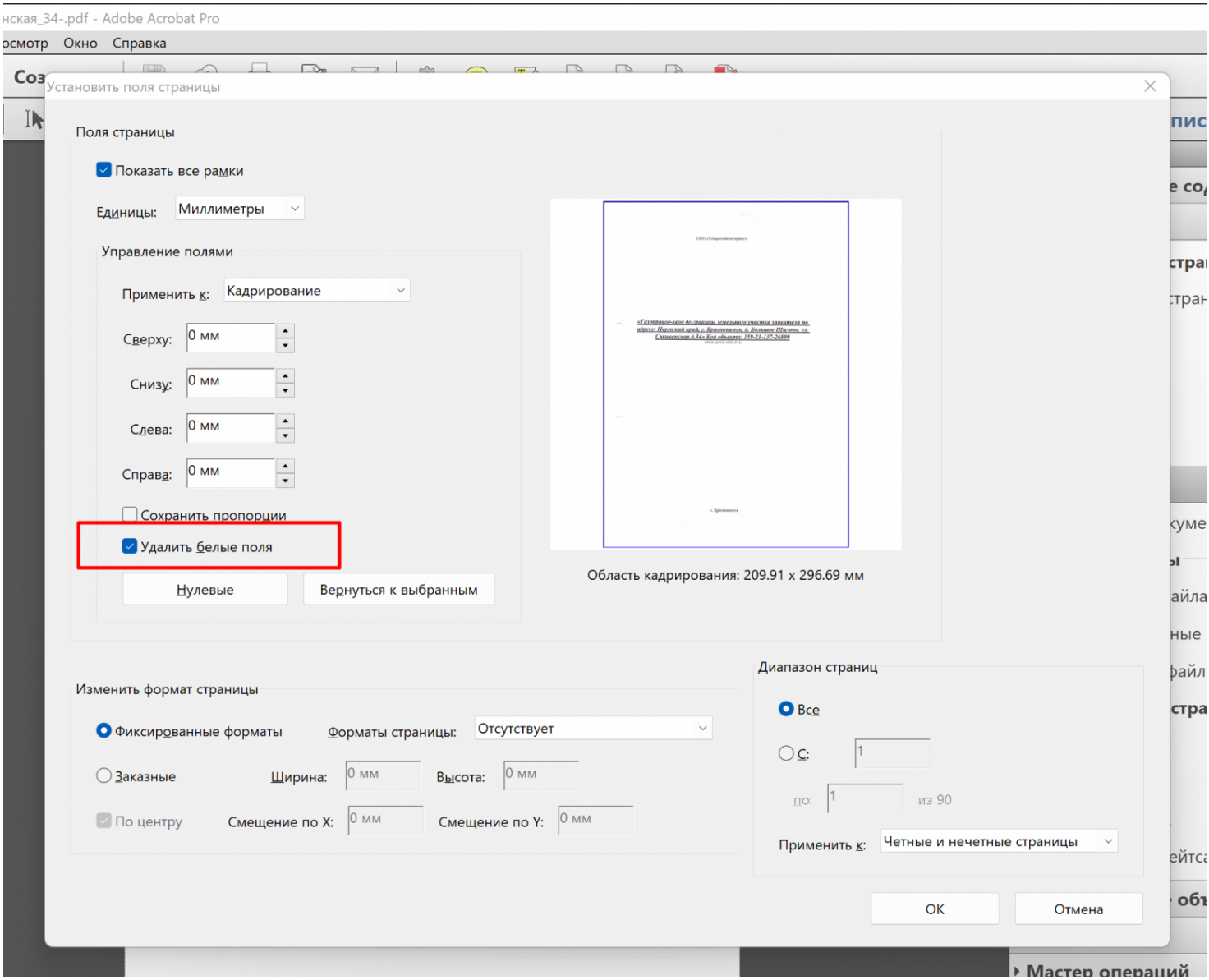
3. Выберите «Страницы» из списка инструментов справа.

4. Нажмите «Обрезать» в меню инструментов «Страницы».

5. В диалоговом окне «Обрезка страниц» выберите параметры «Удалить белые поля» и «Удалить белые поля для всех страниц».

6. Нажмите «ОК», чтобы применить изменения.

   Эти действия должны были автоматически удалить все пустые страницы из файла PDF, но у меня этого не произошло.



Adobe Acrobat Pro и удаление пустых страниц

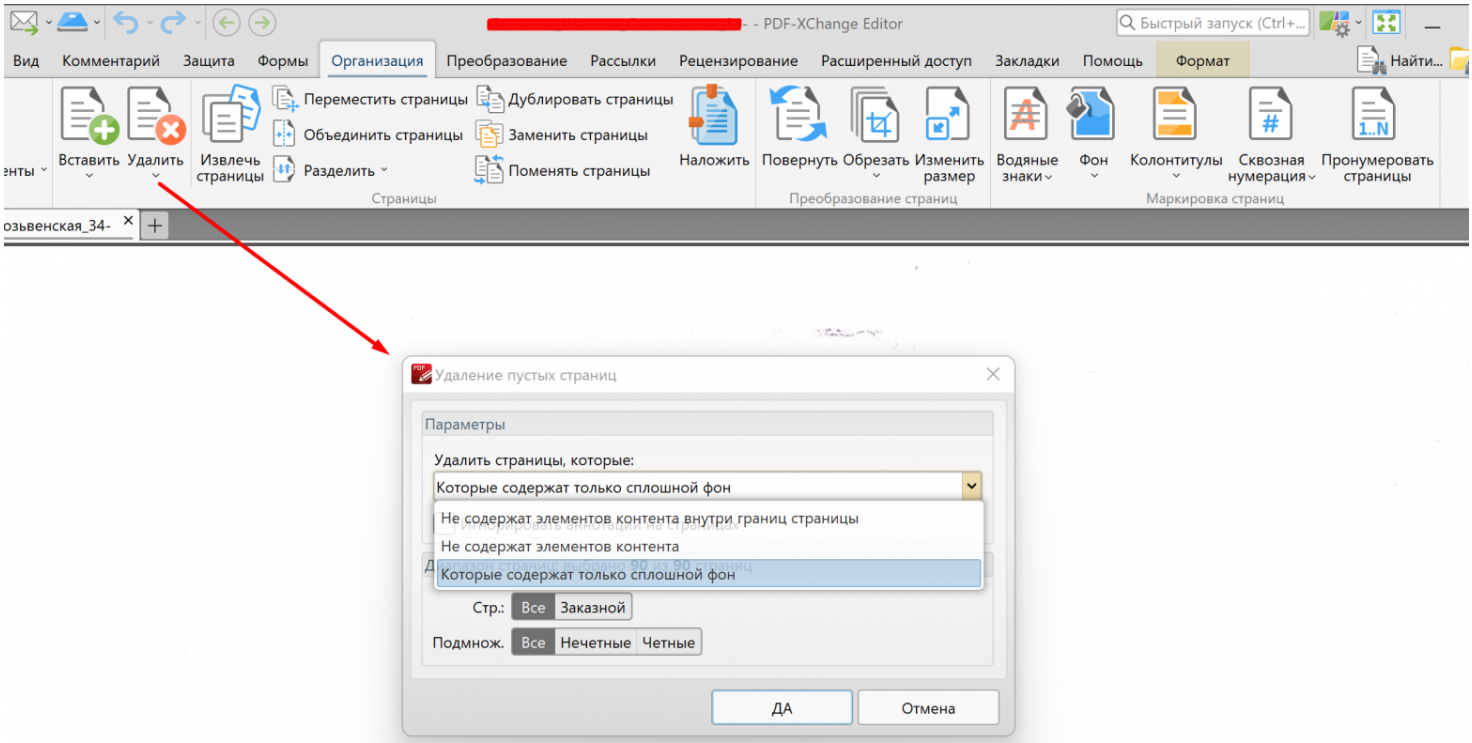
3. Пытался сделать это при помощи PDF-XChange Editor, но у меня тоже не получилось. У меня была инструкция:

   1. Загрузите файл PDF: выберите «Файл» > «Открыть» или нажмите Ctrl + O на

клавиатуре, затем найдите и выберите файл PDF, из которого вы хотите удалить пустые страницы.

2. После загрузки PDF-файла щелкните вкладку «Организация» на верхней панели инструментов.
3. Выбрав все страницы, нажмите кнопку «Удалить пустые страницы».

   Прогресс пробежал, но пустые страницы оставались на месте для любых из трех вариантов.



PDF-XChange Editor

Использование локальной программы, конечно, было бы лучшим вариантом, потому что это гарантировало, что PDF-файлы останутся на компьютере, обеспечивая конфиденциальность и безопасность по сравнению с использованием онлайн-инструментов.

Вариант удаления пустых страниц из pdf при помощи онлайн-инструментов

Но раз с локальными инструментами у меня не пошло, решил попробовать онлайн сервисы.

Я смог найти несколько доступных онлайн-инструментов, которые могли бы помочь автоматически удалить пустые страницы из PDF-файла:

1. Sejda (https://www.sejda.com/delete-pdf-pages)
2. Smallpdf (https://smallpdf.com/delete-pages-from-pdf)
3. DeftPDF (https://deftpdf.com/delete-pdf-pages)

Ни в одном из них я не смог найти опцию автоматического распознавания пустых страниц, хотя в поисковике попадались ссылки на несуществующие сейчас страницы (pdf remove blank pages) этих сервисов.

Ну и конечно использование онлайн-инструментов может поставить под угрозу конфиденциальность и безопасность ваших документов.

Вариант удаления пустых страниц из pdf при помощи локального bash скрипта и консольной программы PDFtk

После постигшей неудачи решил написать свой собственный скрипт который удалит пустые страницы из всех пдф файлов в текущем каталоге.

При изучении вопроса наткнулся на большую дискуссию, где обсуждался вопрос как лучше удалить пустые страницы из pdf при помощи командной строки. Предлагались разные методы, но у меня были все документы сканированные и это значит, что даже на пустом листе какая-то информация всё равно была — сканы отверстий под перешивку или просто грязь со сканера.

Решил что будет следующий алгоритм:

1. Разделяю PDF документ на отдельные файлы.
2. Страницы меньше определенного размера удаляю.
3. Склеиваю оставшиеся страницы обратно.
4. Повторяю столько раз, сколько PDF файлов в текущей папке.
5. PROFIT

После нехитрых манипуляций получился файл `blank_page_remover.sh`:

```
#!/bin/bash
# Убирает пустые страницы из PDF после двухстороннего сканирования
# Описание в статье https://habr.com/ru/articles/733754/
datetime=$(date +"%Y-%m-%d_%H-%M-%S")
# Создаём единый лог файл для всех действий
log_file="blank_page_remover_$datetime.log"
touch $log_file
# Перебираем все PDF файлы в текущем каталоге
```

```
for file in *.pdf; do
  echo "Работаем с $file..." >> "$log_file"
  # Разделяем PDF файл на отдельные страницы
  echo "Разделяем $file на отдельные страницы..." >> "$log_file"
  pdftk "$file" burst output "${file%.*}_pg_%04d.pdf" >> "$log_file" 2>&1
  # Удаляем файлы страниц, размер которых меньше чем XX килобайт
  echo "Удаляем файлы страниц, размер которых меньше чем 70 килобайт..." >> "$log_file"
  for page in "${file%.*}"_pg_*.pdf; do
    size=$(wc -c < "$page")
    if [[ $size -lt 70000 ]]; then
      echo "Удаляем $page (размер: $size байт)..." >> "$log_file"
      rm "$page"
    fi
  done
  # Склеиваем оставшиеся страницы в новый файл
  echo "Склеиваем оставшиеся страницы в новый файл..." >> "$log_file"
  pdftk "${file%.*}"_pg_*.pdf cat output "${file%.*}_без пустых.pdf" compress >> "$log_fil
  # Удаляем временные файлы
  echo -e "Удаляем временные файлы...\n" >> "$log_file"
  rm "${file%.*}"_pg_*.pdf
done
```

Для работы скрипта понадобится PDFtk (сокращение от PDF Toolkit) — это инструмент командной строки для работы с PDF-файлами. Как его установить для разных операционных систем можно узнать в предыдущей статье.

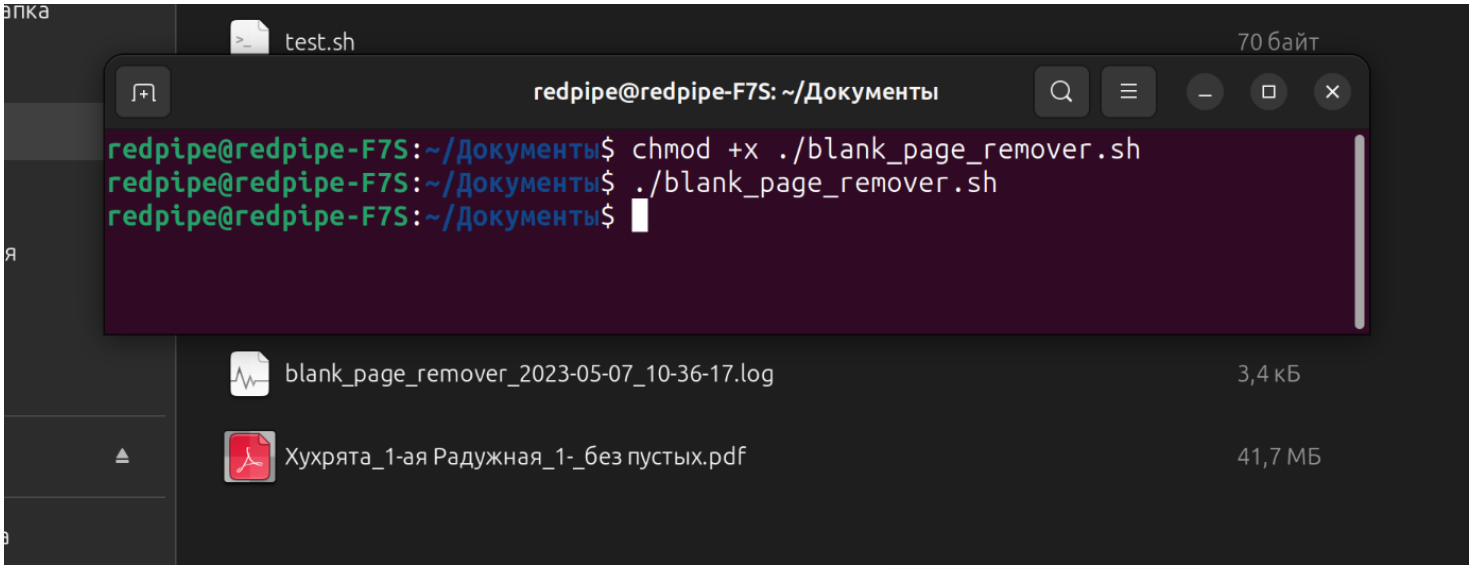
Как воспользоваться скриптом удаления пустых страниц из PDF документа

Чтобы выполнить сценарий bash на компьютере, выполните следующие действия в зависимости от операционной системы:

Для Linux и macOS:

1. Откройте Терминал: нажмите `Ctrl + Alt + T` в Linux или откройте `Терминал` из папки `Приложения > Утилиты` в macOS.
2. Перейдите в каталог, где находится скрипт: используйте команду `cd`, за которой следует путь к каталогу. Например:
   `cd /путь/к/скрипту`
3. Сделайте скрипт исполняемым:
   `chmod +x blank_page_remover.sh`
4. Выполните этот сценарий. Запустите сценарий, введя `./`, а затем имя сценария:
   `./blank_page_remover.sh`

5. PROFIT!
   Скрипт создаст новые пдф файлы без пустых страниц и подробный лог действий.



Терминал в Ubuntu и результат выполнения скрипта `blank_page_remover.sh`

Для Windows (используя GitBash или WSL):

1. Установите GitBash или WSL: если вы еще этого не сделали, установите GitBash или подсистему Windows для Linux (WSL).
2. Откройте Git Bash или WSL: щелкните правой кнопкой мыши папку, содержащую скрипт, и выберите `GitBash здесь` или `Открыть в WSL`.
3. Сделайте скрипт исполняемым:
   `chmod +x blank_page_remover.sh`
4. Выполните этот сценарий. Запустите сценарий, введя `./`, а затем имя сценария:
   `./blank_page_remover.sh`
5. PROFIT!
   Скрипт создаст новые пдф файлы без пустых страниц и подробный лог действий.

Заключение

Удаление пустых страниц из PDF-файлов после двустороннего сканирования может оказаться непростой задачей, особенно при работе с большими объемами документов. Тем не менее, эта статья предоставила вам решение в виде использования автоматического локального сценария bash с консольной программой PDFtk.

Следуя подробным инструкциям вы сможете эффективно избавиться от пустых страниц и поддерживать чистый профессиональный вид отсканированных PDF-документов.

Независимо от объема или сложности ваших файлов, это решение упростит ваш рабочий процесс и сэкономит ваше время и усилия.

Автор: Михаил Шардин,

10 мая 2023 г.

Теги: bash, pdftk, сканирование, документы

Хабы: Open source, PDF, Софт, Лайфхаки для гиков

21 час назад

Убийство разработки: опыт Selectel

8 мин 13K

Обзор

+63 31 18

ru_vds
22 часа назад

Предполётные испытания космического сервера. Спутник «ушёл на золото»

Простой 5 мин 3.3K

+39 22 18

eran
21 час назад

Как мы создаём новые языки в Yandex SpeechKit. Рассказываем на примере узбекского

Простой 8 мин 2.1K

+22 8 7

CyberPaul
20 часов назад

В изоляции. История появления и развития контейнеров

Простой 9 мин 2.2K

Ретроспектива

+21 43 2

Flamyatina
19 часов назад

Доступность сервиса: экспресс-тестирование

Простой 4 мин 714

+20 12 0

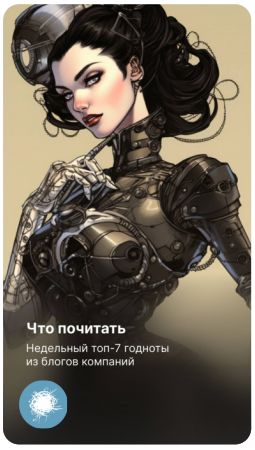
Атлант исправил плечи: вторая линия поддержки IT-проектов ЕВРАЗа

Турбо

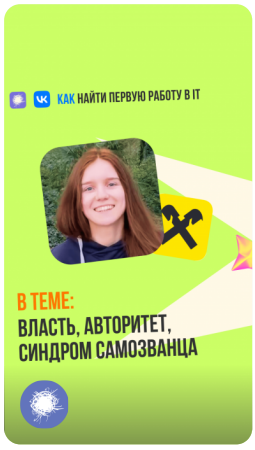
Показать еще

ИСТОРИИ

События для HR и рекрутеров в IT в июне



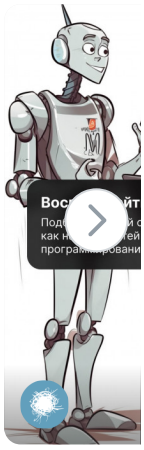
Топ-7 годноты из блогов компаний



Власть, авторитет, синдром самозванца

Чем кулинария похожа на код?

Как взлететь к звёздам с «Промо»



Воспитай айтишника

ВАКАНСИИ

Архитектор инфраструктуры
от 215 000 до 235 000 ₽ · Специальные технологии контроля · Можно удаленно

Администратор Linux
от 140 000 ₽ · Sportmaster Lab · Москва

DevOps/Support/Delivery Manager
от 200 000 до 400 000 ₽ · Сбер · Москва

Администратор тестовых стендов (DevOps)
от 100 000 ₽ · СберТех · Москва

Системный администратор Linux (вторая линия технической поддержки)
от 80 000 до 100 000 ₽ · Support IT · Екатеринбург · Можно удаленно

Больше вакансий на Хабр Карьере

МИНУТОЧКУ ВНИМАНИЯ

Москвичу сложно понять, почему Новгород — нижний, а не правый

Глупым вопросам и ошибкам — быть! IT-менторство на ХК