

Хабр



КАК СТАТЬ АВТОРОМ



Войти



empenoso

10 мая 2023 в 03:42

## Как убрать пустые оборотные страницы из PDF после двухстороннего сканирования

Средний 6 мин 7.9K

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?



Habr в Telegram

Читай только самое важное!



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:

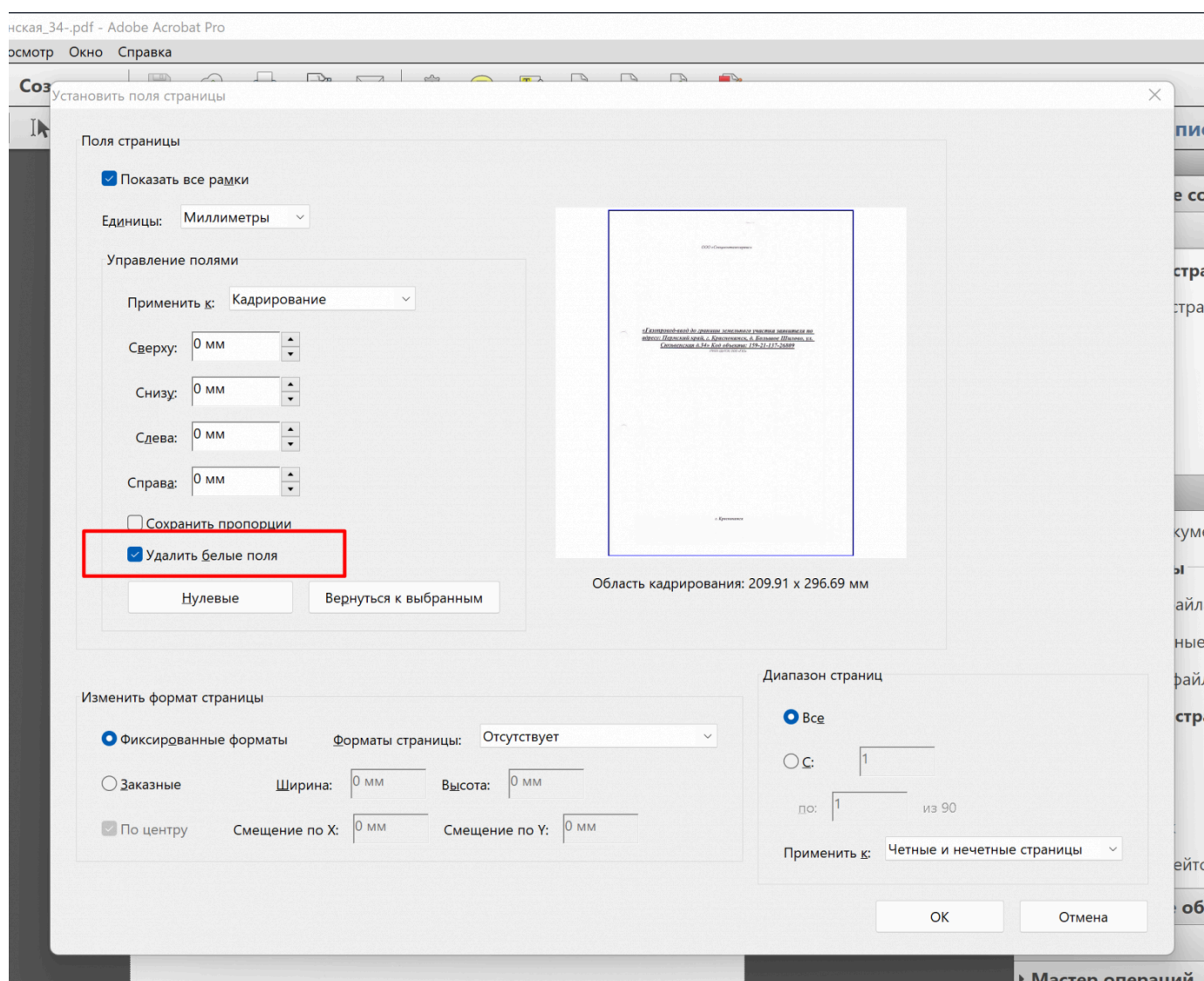
1. Пытался сделать это при помощи бесплатной открытой [PDFsam Basic](#), которая доступна под Linux и Windows, и MacOS, потому что в интернете нашёл инструкции, но они оказались устаревшими.
2. Пытался сделать это при помощи Adobe Acrobat Pro, но у меня не получилось. Делал по инструкции\*

**Habr в Telegram**

Читай только самое важное!

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

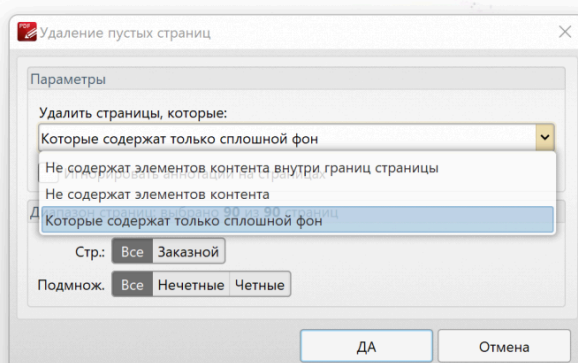
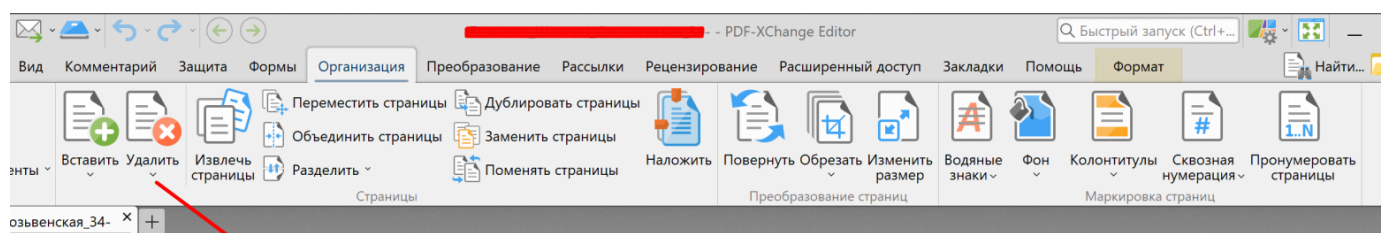


**Habr в Telegram**

Читай только самое важное!

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.



**Habr в Telegram**

Читай только самое важное!



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 pdf 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.



**Habr в Telegram**

Читай только самое важное!

После нехитрых манипуляций получился файл `blank_page_remover.sh` :

```
# Подробнее в статье Как убрать пустые оборотные страницы из PDF после двухстороннего с
# https://habr.com/ru/articles/733754/
# Михаил Шардин https://shardin.name/

#!/bin/bash
datetime=$(date +"%Y-%m-%d_%H-%M-%S")
# Создаём единый лог файл для всех действий и папку куда перемещаем вырезанные страницы
log_file="blank_page_remover_${datetime}.log"
touch $log_file
mkdir removed
# 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" >> "$log_file" 2>&1
    # Удаляем файлы страниц, размер которых меньше чем XX килобайт
    echo "Удаляем файлы страниц, размер которых меньше чем 35 килобайт..." >> "$log_file"
    for page in "${file%.*}_pg_*.pdf; do
        size=$(wc -c < "$page")
        if [[ $size -lt 35000 ]]; then
            echo "Удаляем $page (размер: $size байт)..." >> "$log_file"
            mv "$page" "removed/"
            #rm "$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_
    # 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



**Habr в Telegram**

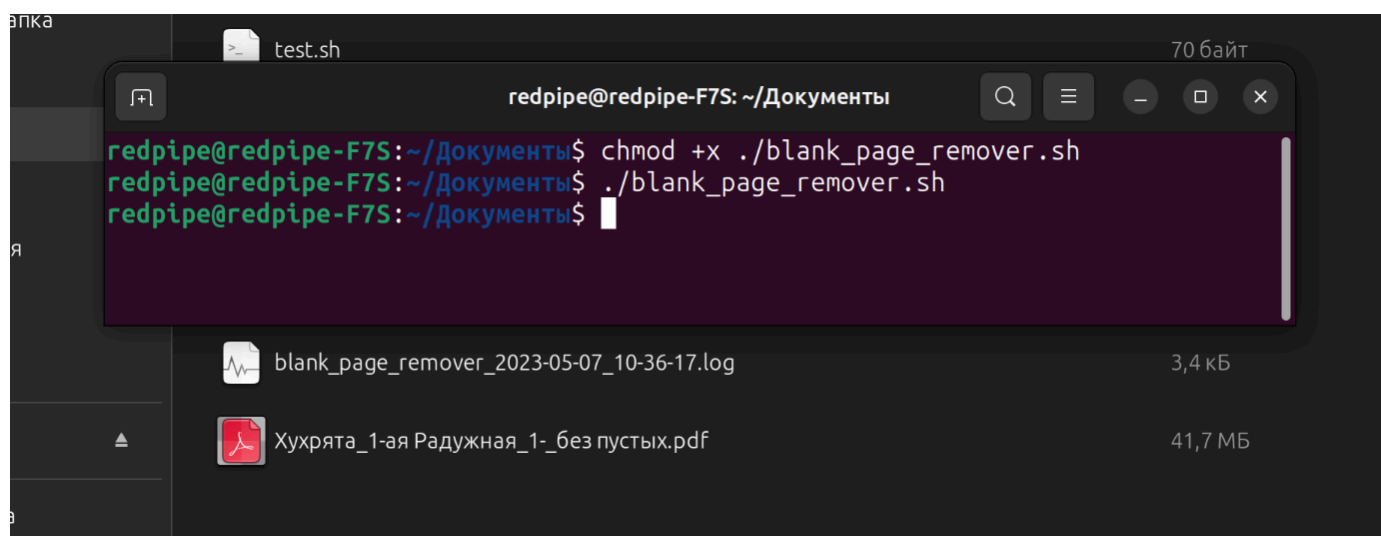
Читай только самое важное!

## 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 pdf 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



**Habr в Telegram**

Читай только самое важное!

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 pdf файлы без пустых страниц и подробный лог действий.

Актуальная версия скрипта [всегда доступна на гитхабе](#).

## 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](#)

## Редакторский дайджест

Присылаем лучшие статьи раз в месяц



**Habr в Telegram**

Читай только самое важное!







156

34.2

Карма

Рейтинг

**Михаил Шардин** @empenoso

Разработчик

[Подписаться](#)[Сайт](#) [Сайт](#) [Github](#)

Комментарии 10

## Публикации

[ЛУЧШИЕ ЗА СУТКИ](#)[ПОХОЖИЕ](#)**Tirarex**

20 часов назад

### Gameboy Advance — полный гайд по выживанию в 2024 году



Простой



11 мин



6.4K

[Тutorial](#)

+39



18



13

**yadro\_team**

23 часа назад

### Как учить языки программирования и создавать базу знаний с помощью метода из прошлого века: опыт четырех инженеров



Простой



10 мин



7.3K

[Обзор](#)**Habr в Telegram**

Читай только самое важное!

**melnik909**

вчера в 14:00

## Магия CSS на практике: советы по вёрстке от гика. Часть 3

**Средний**

6 мин



4.1K

Тutorial

**+26**

57



9

**myops**

19 часов назад

## Удалёнка до того, как стала удалёнкой

**Простой**

22 мин



2.7K

Обзор

**+23**

13



7

**Gradiens**

1 час назад

## А ваша зарплата в рынке? Простой, как топор, способ это узнать

**Простой**

8 мин



3.8K

Мнение

**+17**

15



0

**nike\_ilin**

23 часа назад

## Укрощение ClickHouse: почему ДанКо делает Visiology намного быстрее

**Средний**

10 мин



2.1K

Обзор

**+16**

17



3

**Habr в Telegram**

Читай только самое важное!

## «От идеи и до продакшена»: как разработать веб-приложение и загрузить в VK Mini Apps

 Средний  20 мин  1.9K

Тutorial

 +15

 29

 1



kubelet

3 часа назад

## Kubernetes 1.31: новый VolumeSource, эмуляция старых версий и настройка анонимного доступа к эндпоинтам

 Средний  24 мин  745

Обзор

 +14

 8

 1



Andrevich

12 часов назад

## Решаем проблему блокировок (и YouTube) за 5 минут на роутере с OpenWRT

 Простой  5 мин  18K

Tutorial

 +14

 116

 34



MarinaShmelevaEng

12 часов назад

## Список из 100 полезных фраз для IT на английском языке с примерами употребления

 12 мин  2.1K

Из песочницы

 +14

 42

 6



Habr в Telegram

Читай только самое важное!

## Автоматизировали настройку и установку решения — всё из-за факапа с юзером

[Турбо](#)[Показать еще](#)

### ВАКАНСИИ

#### Прикладной администратор SberApps

от 230 000 ₽ · Сбер · Москва

#### Ведущий администратор Oracle

до 300 000 ₽ · SM Lab · Москва · Можно удаленно

#### Системный администратор Linux

до 200 000 ₽ · Точка · Екатеринбург

#### Эксперт по инфраструктуре

до 200 000 ₽ · Гринатом · Москва

#### DevOps

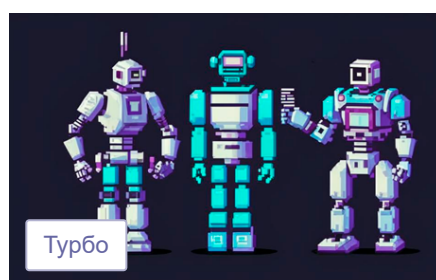
от 120 000 ₽ · N.Work · Можно удаленно

[Больше вакансий на Хабр Карьере](#)

### МИНУТОЧКУ ВНИМАНИЯ



Куда сходить: событийная афиша Хабра



3D-модели по промту и роботы: что было на IT-ивенте GigaConf



Вы сами это читали: как стать лучшим автором месяца



**Habr в Telegram**

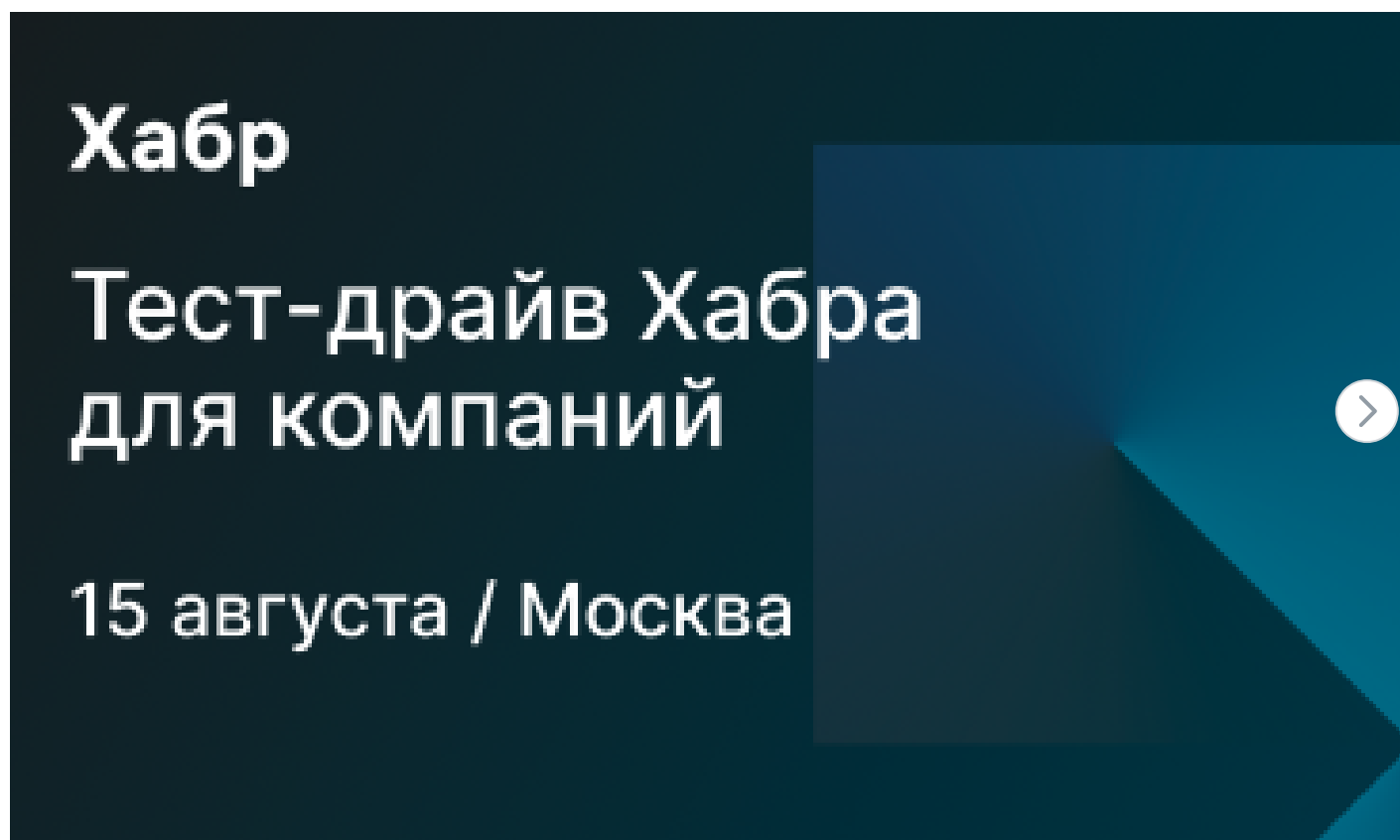
Читай только самое важное!

## БЛИЖАЙШИЕ СОБЫТИЯ

# Хабр

## Тест-драйв Хабра для компаний

15 августа / Москва



15 августа

**Бесплатный тест-драйв Хабра для компаний**

Москва

Другое

[Больше событий в календаре](#)

Хабр

**Habr в Telegram**

Читай только самое важное!



Техническая поддержка

© 2006–2024, Habr



**Habr в Telegram**

Читай только самое важное!