ПДн-Аудитор

веб-сервис аудита документов и веб-ресурсов на утечки персональных данных (152-ФЗ)

Аннотация. Рассматривается разработка веб-сервиса для автоматизированного поиска утечек персональных данных (ПДн) в электронных документах и на веб-ресурсах, а также для приведения документов в соответствие с требованиями Федерального закона № 152-ФЗ «О персональных данных». Сервис объединяет восемь независимых проверок — от извлечения текста из-под фиктивной «замазки» в PDF до анализа политики конфиденциальности сайта с помощью большой языковой модели — в единый отчёт с воспроизводимой оценкой риска. Реализована модульная плагинная архитектура, обеспечивающая расширяемость без изменения ядра, и комплекс мер защиты от типовых атак на сервис, обрабатывающий недоверенный ввод.

Ключевые слова: *персональные данные, 152-ФЗ, обезличивание, утечки данных, аудит документов, редактирование PDF, метаданные, веб-трекеры, защита от промт-инъекций.*

1. Введение и актуальность

Публикация электронных документов и эксплуатация веб-ресурсов сопровождаются регулярными непреднамеренными утечками персональных данных. Причина, как правило, не во внешней атаке, а в незнании того, что файл или страница хранят больше, чем видно на экране.

Наиболее показательный пример — «замазка» в PDF: чёрный прямоугольник, наложенный поверх текста, визуально скрывает данные, но сам текст остаётся в файле и извлекается простым копированием. Аналогично, документы Microsoft Word и PDF сохраняют служебные метаданные (автор, организация, программа, локальные пути), а при включённом режиме рецензирования — удалённые фрагменты и комментарии, невидимые в обычном представлении. Веб-ресурсы, в свою очередь, зачастую нарушают статью 18.1 152-ФЗ: не публикуют политику обработки ПДн и передают идентификаторы посетителей сторонним трекерам без согласия.

Существующие инструменты решают эти задачи разрозненно и ориентированы на технических специалистов. Актуальна разработка единого сервиса, доступного неспециалисту (например, юристу или делопроизводителю) и дающего не только перечень проблем, но и готовое средство их устранения.

2. Цель и задачи

Цель работы — создание веб-сервиса, автоматически выявляющего утечки персональных данных в документах и на веб-ресурсах и приводящего документы в соответствие с 152-ФЗ.

Задачи:

- разработать методы обнаружения скрытых данных в PDF и DOCX (текст под замазкой, метаданные, следы правок);
- реализовать распознавание персональных данных в тексте и обезличивание документа с физическим удалением данных из файла;
- реализовать проверку веб-ресурсов на соответствие 152-ФЗ (трекеры, обязательные сведения, разбор политики);

- спроектировать расширяемую архитектуру и обеспечить защиту сервиса от атак, характерных для обработки недоверенного ввода.

3. Архитектура системы

Система построена по принципу «микроядро — плагины». Каждая проверка представляет собой независимый модуль в каталоге checks/, а неизменяемое ядро отвечает за их обнаружение, запуск и сбор результатов. Архитектура держится на трёх принципах:

- **Единый контракт результата.** Любая проверка возвращает список объектов Finding одной формы (идентификатор, уровень, суть, объяснение, местоположение, доказательство, рекомендация).
- **Реестр плагинов.** Декоратор @register и автодискавери находят все проверки и запускают те, что применимы к типу входа (PDF, DOCX или URL).
- **Изоляция ошибок.** Оркестратор выполняет каждый плагин в отдельном блоке try/except: сбой одной проверки не прерывает остальные, а попадает в отчёт как находка уровня «ошибка».

Фронтенд отображает находки универсальным способом непосредственно из схемы данных, поэтому новый плагин появляется в интерфейсе автоматически. Правило проекта: новая функция — это новый файл в checks/ и одна строка регистрации; ядро после начальной фазы не изменяется. Такой подход обеспечил последовательную реализацию функциональности восемью фазами без нарастания связанности кода.

4. Функциональные возможности

Сервис работает в двух режимах — анализ документа (PDF/DOCX) и анализ веб-ресурса (URL) — и реализует восемь проверок:

Вход	Проверка	Что обнаруживает и как
PDF	Текст под замазкой	Извлекает текст из-под чёрных прямоугольников (движок x-ray, PyMuPDF): «замазка» — фигура поверх, текст остаётся в файле.
PDF / DOCX	Скрытые метаданные	Автор, редактор, организация, ПО, даты, локальные пути (C:\Users\Фамилия\…) из служебных полей файла.
DOCX	Забытые правки	Непринятые вставки/удаления (w:ins/w:del) и комментарии с автором и датой из word/document.xml.
PDF / DOCX	Личные данные	ФИО (Natasha), ИНН, СНИЛС, паспорт, телефон, счёт, адрес (Presidio + правила для РФ).
PDF / DOCX	Обезличивание	Отдаёт очищенную копию: данные удаляются из файла физически, замена по судебному образцу.
URL	Куки и трекеры	Реальная загрузка в Chromium (Playwright), классификация доменов по DuckDuckGo Tracker Radar.
URL	Обязательные сведения	Чек-лист ст. 18.1 152-ФЗ: политика ПДн, реквизиты оператора, согласие, контакты.
URL	Разбор политики	LLM (xAI Grok) со строгой JSON-схемой раскладывает политику по пунктам 152-ФЗ.

Результаты сводятся в единый отчёт: находки ранжируются по уровню (критично, предупреждение, инфо, сделано правильно), вычисляется интегральная оценка риска (высокий/средний/низкий) и формируется экспорт в PDF с поддержкой кириллицы.

5. Обезличивание по судебному образцу

Ключевая практическая функция — подготовка документа к публикации. В отличие от «замазки», сервис удаляет персональные данные из файла физически (механизм редакции PyMuPDF для PDF и правка XML для DOCX), после чего повторная проверка очищенной копии не находит данных.

Условные обозначения соответствуют практике публикации судебных актов: физические лица обозначаются как ФИО1, ФИО2 (сквозная нумерация в порядке появления), прочие сведения заменяются на «данные изъяты», «адрес», «номер телефона». Реализована устойчивость к особенностям русского языка: одному лицу присваивается единый номер, даже если оно упоминается то по фамилии, то по имени-отчеству; сокращения не склоняются, поэтому текст остаётся читаемым в любом падеже. Распознавание ФИО выполняется библиотекой Natasha, прочих идентификаторов — Presidio с правилами для российских форматов; вся обработка происходит офлайн.

6. Информационная безопасность

Поскольку на вход сервиса поступают недоверенные файлы и произвольные адреса, границы защиты заданы явно:

- **Защита от промт-инъекций.** Текст политики со стороннего сайта передаётся языковой модели как данные: применяются строгая JSON-схема вывода (ответ физически ограничен фиксированным набором полей), системный запрет исполнять команды из данных и обёртка недоверенного текста разделителями со случайным одноразовым маркером (nonce). Эффективность подтверждена тестами со встроенной атакой.
- **Защита от SSRF.** Загрузка по пользовательскому URL отсекает не-HTTP(S)-схемы и непубличные адреса (localhost, приватные диапазоны, метаданные облака).
- **Path traversal и лимиты.** Имя загружаемого файла обрезается до базового, запись — только во временный каталог; объём загрузки ограничен, чтение потоковое.
- **Контроль доступа.** Сервис закрыт паролем (сравнение, устойчивое ко времени), перебор блокируется, действует ограничение частоты запросов с одного адреса.

Принцип приватности: данные не хранятся дольше сессии — загруженный файл удаляется сразу после анализа, сгенерированные артефакты очищаются по истечении срока жизни, содержимое с ПДн нигде не журналируется.

7. Технологический стек

Бэкенд — Python 3.11, FastAPI; обработка документов — PyMuPDF и lxml; распознавание ПДн — Presidio и Natasha; анализ веб-ресурсов — Playwright (Chromium); экспорт отчёта — ReportLab. **Фронтенд** — React, Vite, Tailwind CSS. **Развёртывание** — Docker Compose (бэкенд, статика через nginx-прокси) в одну команду. Внешний платный вызов один — разбор политики через API xAI (Grok); без ключа сервис корректно деградирует, сохраняя остальную функциональность.

8. Апробация на реальном документе

Сервис проверен на реальном мотивационном письме объёмом около 6,5 тыс. знаков. Персональные данные распознаны, и возвращена обезличенная копия, оформленная по практике публикации судебных актов: упомянутые лица заменены на ФИО1 и ФИО2 (одному лицу присвоен единый номер, несмотря на упоминания по фамилии, по имени-отчеству и полным ФИО), организации — на «<организация>». Данные удалены из файла физически, служебные метаданные вычищены; всего применено 13 замен. Фрагмент обезличенного документа приведён на рис. 1, карта замен — на рис. 2.

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

Мотивационное письмо для поступления на магистерскую программу «<организация>»

Здравствуйте, уважаемые коллеги!

Меня зовут ФИО1 , я бакалавр юриспруденции МГУ. Большинство абитуриентов программы «<организация>» приходят, чтобы научиться мыслить технически. Я прихожу за другим, потому что уже мыслю так. Я строю системы: вижу задачу целиком, нахожу в ней слабое место и собираю решение. Во время учёбы я писал программы, проектировал устройства и разбирался, как можно применить технологии в праве. «<организация>» стал первой программой, где подобный характер работы становится профессией. В магистратуре я рассчитываю не сменить мышление, а получить инструменты, среду и научное сообщество, которые позволят раскрыть его в полную силу и строить собственные проекты на стыке юриспруденции и технологий.

Моя цель на ближайшие годы состоит не в том, чтобы занять определённую должность, а в том, чтобы научиться строить работающие продукты на стыке права и технологий: проектировать и запускать решения, которые оптимизируют и автоматизируют юридическую функцию. Роль руководителя направления или ведущего эксперта для меня следствие этого, а не самоцель.

Достижения в «Лигал» и «Тех» по раздельности: я занял 1 место в международном конкурсе «Ars Sacra» в секции финансового права, а также первые места в международных научно-исследовательских конкурсах «<организация>» и «Лучшие научные работы 2026» в секции юридических наук. Отдельно отмечу две публикации РИНЦ по прикладному искусственному интеллекту в праве. Это направление я планирую развивать и дальше: в программе ему посвящена дисциплина «Искусственный интеллект и его регулирование», а я рассчитываю сделать его темой магистерской и кандидатской работы.

Являлся членом студенческого совета, кандидатом на вступление в СНО факультета, был старостой и куратором младших курсов. Прошёл военную подготовку в <организация> МГУ, являюсь сержантом запаса.

Рисунок 1 — Фрагмент реального документа после обезличивания сервисом

✓ Документ обезличен

Заменено фрагментов: **13**. Данные удалены из файла, а не закрашены — под заменой ничего не осталось. Метаданные вычищены.

Что на что заменили

ФИО	~~Мамонов Николай Павлович~~	→ **ФИО1**
ФИО	~~Лагутин Максим Дмитриевич~~	→ **ФИО2**
организация	~~Военный учебный центр~~	→ **<организация>**
организация	~~ЛигалТех~~	→ **<организация>**
организация	~~Научный олимп~~	→ **<организация>**

Рисунок 2 — Карта замен, выполненных сервисом

В ходе апробации выявлены и устранены две неточности распознавания. Во-первых, модуль принимал за организацию сокращение «ПДн» (аббревиатура «персональные данные») и обрывок «Тех», отделённый сегментатором от названия «ЛигалТех»: сокращение смешанного регистра не опознавалось как аббревиатура, а короткий обрывок не отсеивался. Во-вторых, имя, которое текстовый редактор хранит несколькими фрагментами, давало сдвоенный псевдоним вида «ФИО1 ФИО1». Обе неточности устранены точечной правкой соответствующих модулей, дополнены регрессионными тестами и не затронули ядро и остальные проверки. Это подтверждает практическую ценность выбранной архитектуры: доработка локализуется в отдельном модуле.

9. Заключение

Разработан работоспособный веб-сервис, объединяющий восемь проверок документов и веб-ресурсов на утечки персональных данных в единый отчёт с оценкой риска и предоставляющий готовое средство обезличивания. Модульная плагинная архитектура обеспечивает расширяемость без изменения ядра, а реализованный комплекс мер защищает сервис от типовых атак на обработчик недоверенного ввода. Практическая значимость состоит в том, что инструмент делает доступной неспециалисту проверку, обычно требующую технической квалификации, и тем самым снижает риск непреднамеренного разглашения персональных данных при публикации документов и эксплуатации веб-ресурсов.

Приложение. Экраны сервиса

Снимки экрана рабочего интерфейса сервиса.

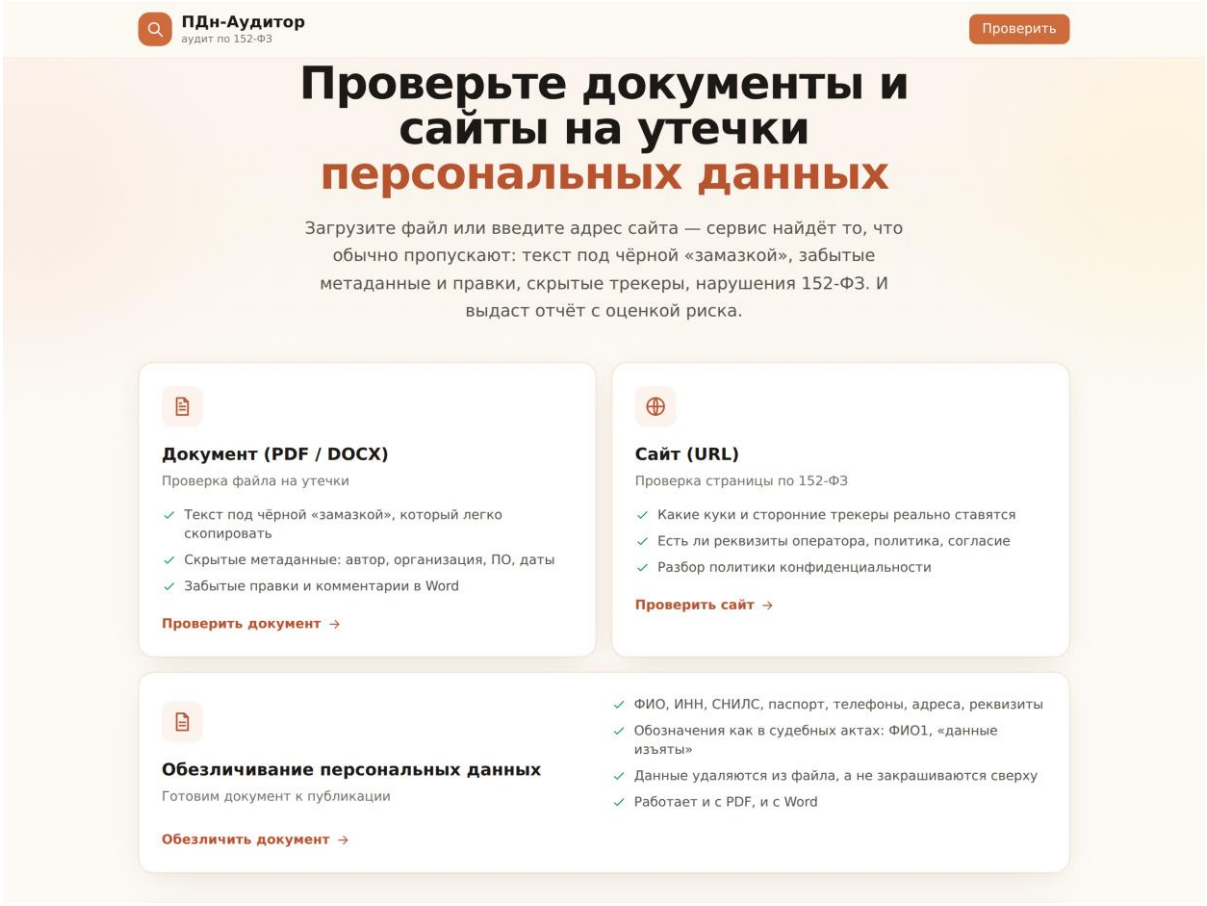

Рисунок 3 — Главный экран: назначение сервиса и три режима работы

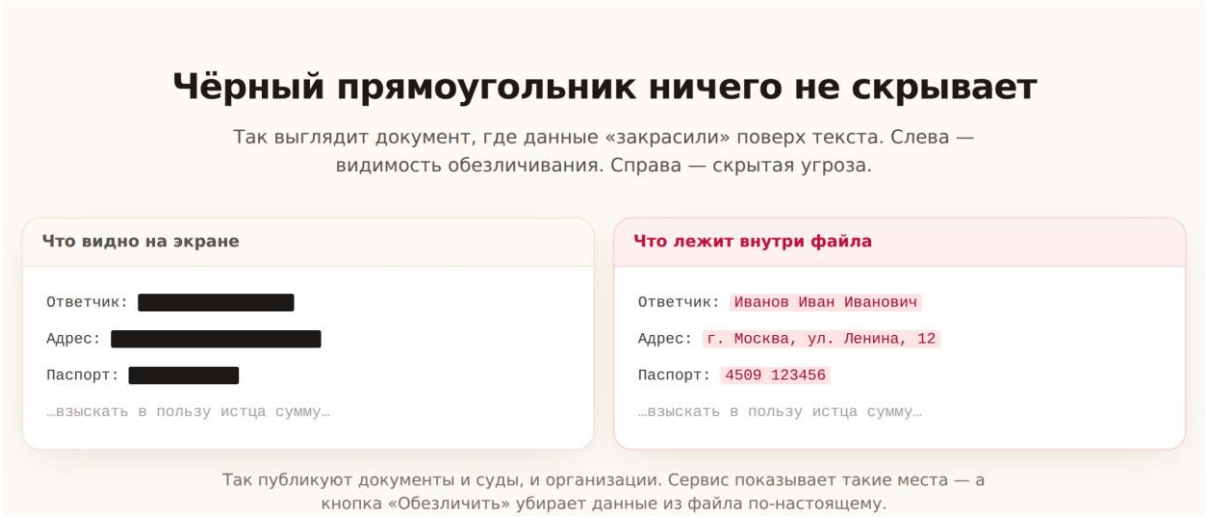

Рисунок 4 — Проблема «замазки»: слева — вид на экране, справа — содержимое файла

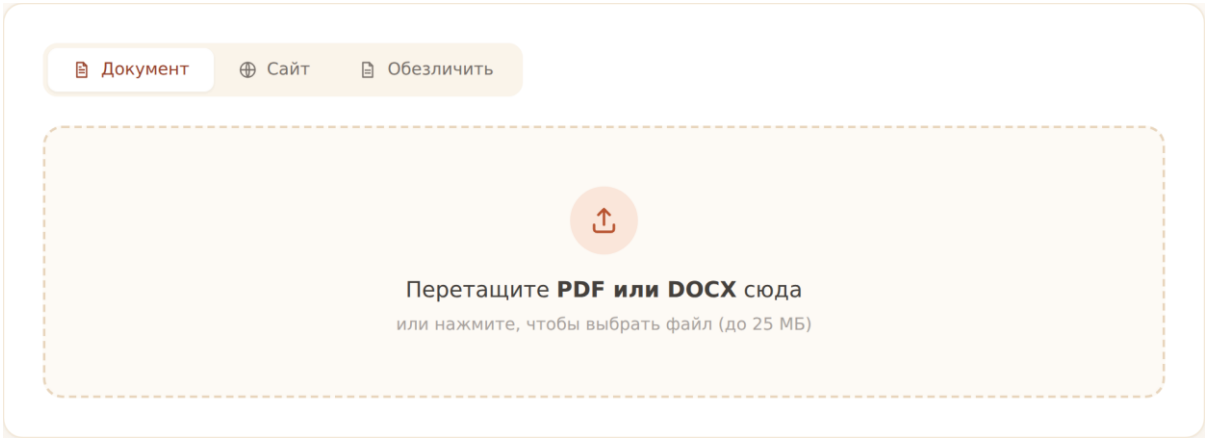

Рисунок 5 — Загрузка документа для проверки

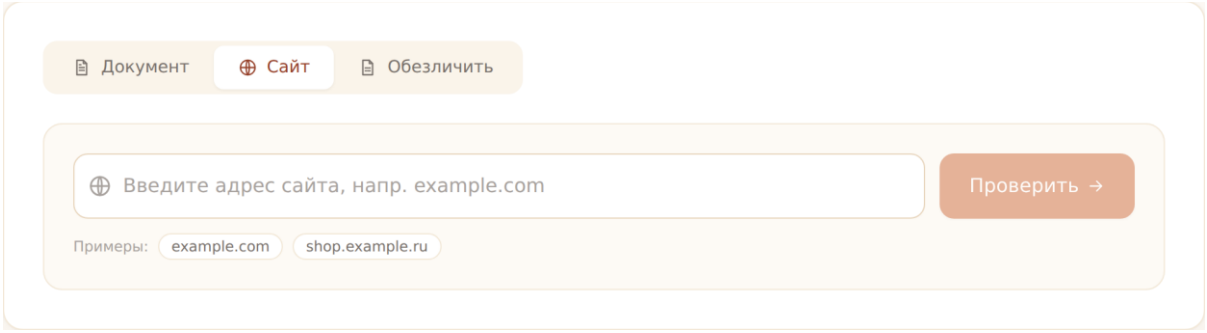

Рисунок 6 — Проверка веб-ресурса по адресу

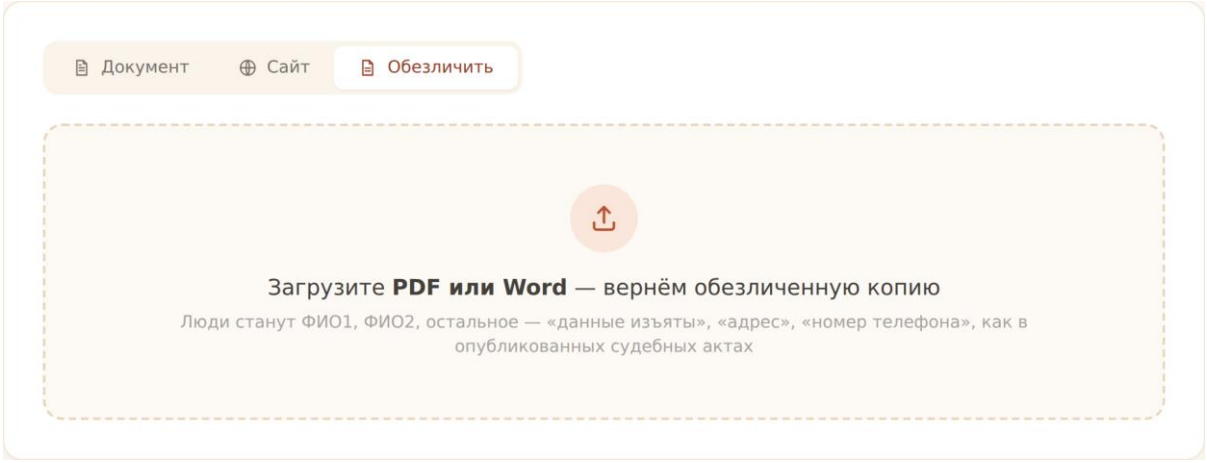

Рисунок 7 — Обезличивание документа

Уровень риска: высокий критично: 1 предупреждений: 2 сделано хорошо: 1

Скачать отчёт PDF

Найдены критичные проблемы — данные утекают прямо сейчас.

Отчёт по: Договор_оказания_услуг_2024.pdf · находок: 4

✓ **Готова очищенная копия документа**

Персональные данные удалены физически, а не закрашены. Можете скачать и проверить повторной загрузкой — находок быть не должно.

Скачать обезличенную копию

Текст под чёрной «замазкой» читается Критично

В документе есть чёрные прямоугольники, но текст под ними остался в файле целиком. Его можно выделить мышью, скопировать и найти поиском — «замазка» только визуальная.

Где: стр. 2

Что нашли: Сидоров Пётр Николаевич, паспорт 45 09 № 123456, оклад 180 000 ₽

Рекомендация: **Нажмите «Обезличить» — сервис удалит текст физически, а не закрасит сверху**

Проверка: redaction

В тексте открытые персональные данные Предупреждение

Найдены сведения, по которым человека можно однозначно опознать. Перед публикацией документ нужно обезличить.

Где: стр. 1, стр. 3

Что нашли: ФИО — 3, ИНН — 1, телефон — 2, адрес — 1, номер счёта — 1

Рекомендация: **Нажмите «Обезличить» — данные заменятся по судебному образцу (ФИО1, «данные изъяты»)**

Проверка: anonymize

В свойствах файла раскрыты автор и организация Предупреждение

PDF хранит служебные поля, которых не видно в самом тексте: кто создал документ, в какой программе и когда. Туда же попал локальный путь с фамилией.

Где: поля Author, Producer

Что нашли: Author: Иванова М. А. · Company: ООО «Правовой центр» · C:\Users\ivanova\Договоры\

Рекомендация: **Нажмите «Обезличить» — служебные поля будут очищены**

Проверка: metadata

Подпись и печать оформлены корректно Хорошо

Графические элементы подписи не содержат извлекаемого текста с персональными данными.

Где: стр. 4

Проверка: redaction

Рисунок 8 — Отчёт по документу: находки и интегральная оценка риска

Уровень риска: высокий критично: 1 предупреждений: 2 сделано хорошо: 1

Скачать отчёт PDF

Найдены критичные проблемы — данные утекают прямо сейчас.

Отчёт по: medcentr-example.ru · находок: 4

На сайте нет политики обработки персональных данных Критично

Статья 18.1 152-ФЗ обязывает оператора опубликовать документ, определяющий политику обработки ПДн, и обеспечить к нему доступ. Ссылку на такой документ найти не удалось.

Где: футер, страница /privacy

Что нашли: Ссылок вида «Политика конфиденциальности / обработки ПДн» на сайте не обнаружено

Рекомендация: Опубликуйте политику обработки ПДн и разместите ссылку в подвале сайта

Проверка: site_disclosures

Сторонние трекеры собирают данные посетителей Предупреждение

Страница при открытии сама отправляет запросы на чужие серверы и ставит куки счётчиков. На чужую сторону уходит IP и идентификатор посетителя — это передача данных третьему лицу.

Где: при загрузке главной страницы

Что нашли: Яндекс.Метрика (mc.yandex.ru), VK Pixel (vk.com), Google Analytics (google-analytics.com)

Рекомендация: Отразите передачу данных этим сервисам в политике и берите согласие до их загрузки

Проверка: cookies

В политике не указан срок хранения данных Предупреждение

Текст политики разобран по пунктам 152-ФЗ. Оператор, цели обработки и права субъекта описаны, но срок хранения персональных данных не назван.

Где: страница /policy

Что нашли: Разделы найдены: оператор, категории данных, цели, права субъекта. Не найден: срок хранения.

Рекомендация: Добавьте в политику срок хранения ПДн или условие прекращения обработки

Проверка: policy

Реквизиты оператора указаны Хорошо

На сайте есть наименование организации и ИНН/ОГРН — субъект понимает, кто обрабатывает его данные.

Где: футер

Проверка: site_disclosures

Рисунок 9 — Отчёт по сайту: проверка соответствия 152-ФЗ

Как это работает

Три шага, никакой настройки. Всё считается на сервере, файлы удаляются сразу после проверки.

1. Вы даёте файл или адрес

Перетаскиваете PDF/Word или вводите адрес сайта. Ничего настраивать не нужно.

2. Сервис проверяет

Документ разбирается по слоям, сайт открывается специальным браузером.

3. Вы получаете результат

Список находок с оценкой риска, обезличенная копия документа и отчёт в PDF — можно приложить к работе.

Что именно проверяем

Восемь проверок. Каждая — отдельный пункт в отчёте: что нашли, почему это важно и что делать.

Документ · Критично

Текст под замазкой

Чёрный прямоугольник в PDF — это фигура, нарисованная поверх. Текст под ней остаётся в файле целиком: выделяется мышью, копируется, находится поиском. Достаём его и показываем вам — ровно тем же способом, каким его достанет посторонний.

Документ · Предупреждение

Кто и где готовил файл

Word и PDF сами дописывают служебные поля: автор, последний редактор, организация, программа, время создания и правок. Туда же попадает путь вроде C:\Users\Фамилия\Договоры. В тексте документа этого не видно — только в свойствах файла.

Word · Предупреждение

Следы правок

При включённом рецензировании удалённый текст не исчезает: он помечается тегом w:del и лежит в document.xml вместе с именем правившего и датой. «Не отображать исправления» убирает их только с экрана. Собираем удалённые фрагменты, примечания и список авторов.

Документ · Предупреждение

Личные данные в тексте

Ищем то, по чему человека можно опознать: ФИО, паспорт, ИНН, СНИЛС, телефон, адрес, счёт, почту. Опасны не только прямые идентификаторы — фамилия рядом с адресом или номером дела указывает на человека не хуже паспорта.

Обезличивание · Даём файл

Замена по судебному образцу

Обозначения взяты из опубликованных судебных актов: люди — ФИО1, ФИО2, остальное — «данные изъяты», «адрес», «номер телефона». Сокращение не склоняется, поэтому фраза читается в любом падеже. Один человек получает один номер во всём документе, даже если его называют то по фамилии, то по имени-отчеству.

Сайт · Предупреждение

Кто следит за посетителем

Открываем страницу настоящим браузером и записываем каждый исходящий запрос и каждую куку. Домены сверяем со списком трекеров (выжимка из DuckDuckGo Tracker Radar): счётчики и рекламные пиксели отправляют на чужие серверы IP и идентификатор посетителя — это передача данных третьему лицу.

Сайт · Критично

Чего не хватает по закону

Ст. 18.1 152-ФЗ обязывает оператора опубликовать политику обработки ПДн. Проверяем четыре вещи: ссылку на политику, реквизиты оператора (ИНН/ОГРН/наименование), формулировку согласия у форм сбора данных и контакт для обращений. Сам текст политики разбираем отдельно.

Итог · PDF

Отчёт на руки

Все находки одним списком: критичные, предупреждения и то, что сделано правильно. У каждой — что нашли, где именно и что с этим делать. Одной кнопкой выгружается в PDF.

Рисунок 10 — Принцип работы и восемь реализованных проверок