

DOCX anomaly inspector

Спецификация реализованного модуля для поиска и безопасной корректировки явных аномалий DOCX. Модуль относится к разделу **СТИЛИ И ГИГИЕНА** и вынесен в отдельную вкладку **АНОМАЛИИ ДОКУМЕНТА**, чтобы не смешивать низкоуровневую текстовую гигиену с анализом структуры документа.

Назначение

DOCX anomaly inspector ищет признаки дурного форматирования, которые обычно видны человеку при вычитке документа: съехавшие таблицы, странные секции, разрывы, пропуски нумерации, оторванные подписи, неоднородные абзацы и похожие системные сбои.

Главный принцип: модуль не должен ругать легитимную сложность документа. Например, альбомная ориентация сама по себе не аномалия, если она нужна для широкой таблицы. Аномалия - альбомная страница с обычным текстом без видимой причины, либо разорванная таблица, где часть оказалась в другой ориентации.

Граница ответственности: этот модуль отвечает за поиск и гигиену ошибок форматирования и структуры. Он не должен становиться общей текстовой вычиткой: двойные пробелы, единицы измерения, адресные обозначения, населённые пункты, улицы, дома и морфологические правила относятся к общей команде **Глубокая гигиена DOCX**, описанной в `docs\DOCX_DEEP_HYGIENE_RU.md`.

Команды UI

Текущий вариант в **СТИЛИ И ГИГИЕНА**:

- Проверка аномалий DOCX** - только отчёт, без изменения файлов. RO-слой реализован в `system_core\docx_anomaly_inspector.py` и покрывает таблицы, подписи таблиц/рисунков, секции/ориентацию, колонтитулы и нумерацию страниц, поля/оглавление/ссылки/закладки, списки/нумерацию, пустоты и заголовки.

- **Корректировка аномалий DOCX** - применяет только выбранные безопасные исправления и пишет отчёт до/после. Первый слой реализован в `system_core\docx_anomaly_corrector.py`: лишние пустые абзацы, точная высота строк, запрет переноса текста в ячейках, опциональная унификация границ и полей ячеек.

В UI это выглядит как одна вкладка **АНОМАЛИИ ДОКУМЕНТА** с переключателем режима:

- **Проверка** - тёмно-зелёный кант, режим без изменения документов.
- **Корректировка** - тёмно-оранжевый кант, режим вмешательства в копии DOCX.

UI показывает чекбоксы классов проверок в адаптивной сетке: подпись стоит на линии чекбокса, а число колонок уменьшается при узкой ширине панели. Кнопка запуска находится в верхней строке справа, на уровне **НАЗАД**. Корректировка всегда пишет копии в **output**, исходники не меняются.

Формат отчёта

Отчёт должен быть табличным и пригодным для ручной проверки:

Поле	Значение
severity	<code>critical</code> , <code>warning</code> , <code>info</code>
class	класс аномалии
object	абзац, таблица, рисунок, секция, поле, закладка
location	человекочитаемая локация: файл, страница, номер раздела, номер таблицы/рисунка
evidence	короткое доказательство, что найдено
why_it_matters	почему это похоже на ошибку
suggested_fix	что можно сделать
autofix	<code>none</code> , <code>safe</code> , <code>review_required</code>

Отчёт должен идти в `report\docx_anomalies.md/.docx/.json`. Копии с исправлениями - в `output\docx_anomaly_fixed`.

Локация ошибок

Основная локация в отчётах должна быть устроена так, как человек реально ищет место в Word: по странице, номеру раздела, номеру таблицы и номеру рисунка. Внутренние индексы абзацев, run-ов и XML-узлов можно писать в JSON как служебные debug-поля, но не использовать как главный ориентир в MD/DOCX отчётах.

Приоритеты локации:

1. **страница** - точная, если доступна через Word COM/rendered pagination; иначе **страница: не определена**.
2. **раздел** - ближайший заголовок с номером, например **2.1.3 Исходные данные**.
3. **объект** - **Таблица 7**, **Рисунок 4**, **Приложение А**, **колонтитул**, **оглавление**, **закладка**.
4. **деталь** - строка/колонок таблицы, подпись, шапка, поле **SEQ**, номер страницы, пустой абзац перед/после объекта.
5. **поисковый фрагмент** - короткая строка, которую можно вставить в поиск Word, если номер страницы или объекта недостаточен.

Примеры:

- **стр. 27; раздел 2.1.3 "Исходные данные"; Таблица 12; строка 4**.
- **стр. 44; Рисунок 7; подпись на следующей странице**.
- **стр. 1; колонтитул; номер страницы начинается с 3**.
- **страница: не определена; раздел 4.2; Таблица 18; шире полезной области**.

Для DOCX без доступной пагинации отчёт всё равно должен давать пригодный человеческий адрес: ближайший раздел, номер таблицы/рисунка/подписи и поисковый фрагмент.

Глоссарий

- Аномалия - форматный или структурный признак, который статистически похож на ошибку в конкретном документе.
- Разрыв секции - служебный объект Word, который меняет поля, ориентацию, колонтитулы или нумерацию страниц.
- Строки перед шапкой (**pre-header**) - строка перед основной шапкой таблицы. Часто это строка с номерами колонок. Если она есть, повторяться при переносе обычно должна именно она, а не весь большой заголовочный блок.

- Фиксированная высота строки - настройка строки таблицы, при которой Word может обрезать переносимый текст.
- Висячий заголовок - заголовок в конце страницы без текста, таблицы или рисунка сразу после него.
- Оторванная подпись - подпись таблицы/рисунка находится далеко от объекта или уехала на другую страницу.

Классы проверок

Типографика и абзацы

Проверять:

- резкий разброс шрифтов и кеглей внутри похожих абзацев;
- разные интервалы и отступы в похожих абзацах;
- случайные жирность, курсив или подчёркивание только в рамках одного слова, заголовка, наименования или названия;
- слишком маленький кегль в основном тексте и таблицах, если это выбивается из документа.

Не считать аномалией:

- легитимный курсив/жирность в длинных цитатах, терминах, названиях, формулах и заголовках;
- цветной текст вне таблиц/заголовков - это уже покрывается существующей гигиеной цвета и подсветки.

Безопасные правки:

- нормализация интервалов/отступов только для групп абзацев, которые уверенно похожи по стилю и роли;
- снятие случайного начертания только при строгом локальном паттерне.

Секции, страницы и ориентация

Проверять:

- разные поля в секциях без видимой причины;
- альбомные страницы с обычным текстом без широкой таблицы или рисунка;
- разорванные таблицы, где части попали в разные ориентации или поля;

- таблицы в портретной секции, если их ширина явно требует альбомной;
- таблицы шириной 200 мм и меньше в альбомной ориентации, когда вокруг нет другой причины для landscape;
- съехавшие колонтитулы;
- разные настройки нумерации страниц, особенно если нумерация уезжает на первый лист;
- пустые и почти пустые страницы.

Не считать аномалией:

- смещение книжной и альбомной ориентации само по себе. Это нормально, если альбомная ориентация нужна широкой таблице.

Безопасные правки:

- только отчёт по секциям и ориентации на первом этапе;
- автоисправление полей/ориентации разрешать позже, отдельным чекбоксом и только в копиях.

Таблицы

Проверять:

- таблица шире полезной области страницы или обрезана;
- таблица не отцентрирована/не вписана в рамках текущей ориентации;
- аномальная ширина таблицы относительно ориентации страницы;
- разные толщины границ внутри одной таблицы;
- разные поля ячеек внутри одной таблицы;
- смесь шрифтов и кеглей внутри одной таблицы;
- слишком плотные ячейки и аномальная балансировка колонок;
- полностью пустые строки или полностью пустые колонки;
- фиксированная высота строк, особенно вместе с отключённым переносом текста;
- отсутствие повторяемой шапки у многостраничной таблицы;
- подозрительные объединения ячеек.

Подозрительные объединения ячеек определять только эвристически:

- единичное объединение в теле таблицы, когда рядом все строки имеют обычную сетку;
- объединение, которое ломает стабильное число колонок в блоке однотипных строк;

- вертикальное объединение с пустым продолжением, из-за которого текст визуально относится не к той строке;
- объединённая ячейка в середине числовых/перечневых данных без похожих объединений выше или ниже;
- объединение в шапке не считать ошибкой по умолчанию.

Правило повторяемой шапки:

- если есть строки перед шапкой с номерами колонок, обычно повторяются только они;
- если строк перед шапкой нет, повторяется распознанная шапка;
- не включать весь большой заголовочный блок в повторяемую шапку без явного сигнала.

Безопасные правки:

- унификация границ;
- унификация полей ячеек;
- снятие фиксированной высоты строк;
- включение переноса текста;
- повтор шапки по строгой эвристике;
- балансировка ширин только через уже существующую таблицную логику DocFlow.

Подписи таблиц и рисунков

Проверять:

- пропуски в нумерации **Таблица N** и **Рисунок N**;
- дубли номеров;
- реверсивность номеров, когда номер уменьшается без перехода в новый логический раздел;
- лишние, недостаточные или нестабильные точки в нумерации;
- пропущенные логические номера;
- подпись далеко от объекта;
- подпись и объект на разных страницах;
- таблица без подписи среди подписанных таблиц;
- рисунок без подписи среди подписанных рисунков;
- подпись после таблицы, когда в документе принято до таблицы, или наоборот.

Нумерация разделов вида **1.** -> **1.1.** -> **1.1.1**, **1.1.2** -> **2.** считается нормальной.

Ошибка - не сама иерархия, а логический пропуск, обратный ход, дубль, нестабильный формат

или несогласованность с соседними объектами.

Безопасные правки:

- отчёт на первом этапе;
- автокоррекция номеров только после отдельного подтверждения, потому что номера могут быть связаны с полями **SEQ** и перекрёстными ссылками.

Оглавление, поля и ссылки

Проверять:

- в оглавлении есть заголовки, которых уже нет в документе;
- заголовки документа не попали в оглавление;
- поля **REF**, **SEQ**, **PAGE**, **PAGEREF** имеют подозрительно несовпадающий отображаемый текст;
- битые ссылки;
- битые или осиротевшие закладки.

Осторожно:

- признак «оглавление устарело» не должен быть громким FAIL без доказательства. Лучше показывать как **warning** с доказательством: расхождение заголовков, страниц или ссылок.

Безопасные правки:

- отчёт и рекомендация обновить поля в Word;
- автообновление полей через Word COM возможно отдельной командой, если Word установлен и пользователь выбрал такой режим.

Нумерация и списки

Проверять:

- сброс нумерации не там;
- пропуски пунктов;
- ручные номера вместо встроенной нумерации списков Word;
- разные стили списков в одном логическом блоке;
- висячие подпункты без родителя;
- лишние или недостающие точки в номерах;

- обратный ход номера внутри одного списка.

Безопасные правки:

- отчёт по умолчанию;
- автоисправление только для очевидных ручных номеров и только после отдельного чекбокса.

Разрывы, пустоты и привязка объектов

Проверять:

- пустые страницы;
- почти пустые страницы: один знак препинания - аномалия, одно слово - вопрос;
- две и более пустые строки подряд;
- разрывы страниц/секций в странных местах;
- заголовок в конце страницы без последующего содержимого;
- таблица оторвана от вводного абзаца или подписи;
- рисунок оторван от подписи.

Не считать аномалией:

- длинные абзацы сами по себе. В документах бывают цитаты, выгрузки, координаты и технические фрагменты, где длина абзаца нормальна.

Безопасные правки:

- удаление лишних пустых абзацев по строгому правилу;
- перенос заголовка вместе со следующим блоком через `keepWithNext`, если это не ломает секции;
- осторожная чистка разрывов только в режиме корректировки.

Матрица автокоррекции

Класс	Автоправка по умолчанию	Причина
Две пустые строки подряд	безопасно	Обычно это след ручной верстки
Фиксированная высота строк	безопасно	Часто режет текст, снятие безопасно

Класс	Автоправка по умолчанию	Причина
Табличные границы/поля ячеек	безопасно	Уже есть проверенная логика унификации
Смесь шрифтов/кеглей в таблице	безопасно/проверить	Безопасно внутри выбранной таблицей команды
Повтор шапки	требует проверки	Нужна осторожная эвристика строк перед шапкой
Ориентация секции	требует проверки	Можно сломать разметку
Нумерация таблиц/рисунков	требует проверки	Может быть связана с полями
Оглавление/REF/PAGE/PAGEREF	требует проверки	Лучше обновлять через Word COM
Битые ссылки/закладки	report_only	Требуется смысловая проверка

UI-рекомендации

- В **Проверка аномалий DOCX** использовать тёмно-зелёный кант: это режим без изменения текста и структуры.
- В **Корректировка аномалий DOCX** использовать тёмно-оранжевый кант: это режим вмешательства в копию документа.
- Все сложные термины должны иметь подсказку: **разрыв секции**, **Строки перед шапкой**, **фиксированная высота**, **PAGEREF**, **SEQ**, **keepWithNext**, **встроенная нумерация Word**, **верхний колонтитул**, **нижний колонтитул**.
- В видимой форме показывать короткие группы параметров, а подробные объяснения переносить в подсказки.

Приоритет реализации

1. Read-only отчёт: секции, ориентация, поля, таблицы, пустоты, подписи, нумерация таблиц/рисунков.
2. Табличные безопасные правки: границы, поля ячеек, фиксированная высота, перенос, повторяемая шапка по строгой эвристике.
3. Абзацные безопасные правки: две пустые строки, похожие абзацы с разными интервалами/отступами.

4. Ссылки/оглавление/поля Word: отчёт и отдельная интеграция с Word COM.